PaperOrchestra: A Multi-Agent Framework for Automated AI Research Paper Writing

Yiwen Song, Yale Song, Tomas Pfister, Jinsung Yoon
Google Cloud AI Research
{yiwensong,yalesong,tpfister,jinsungyoon}@google.com

Abstract

Synthesizing unstructured research materials into manuscripts is an essential yet under-explored challenge in AI-driven scientific discovery. Existing autonomous writers are rigidly coupled to specific experimental pipelines, and produce superficial literature reviews. We introduce PAPERORCHESTRA, a multi-agent framework for automated AI research paper writing. It flexibly transforms unconstrained pre-writing materials into submission-ready LaTeX manuscripts, including comprehensive literature synthesis and generated visuals, such as plots and conceptual diagrams. To evaluate performance, we present PaperWritingBench, the first standardized benchmark of reverse-engineered raw materials from 200 top-tier AI conference papers, alongside a comprehensive suite of automated evaluators. In side-by-side human evaluations, PAPERORCHESTRA significantly outperforms autonomous baselines, achieving an absolute win rate margin of **50%–68%** in literature review quality, and **14%–38%** in overall manuscript quality.

Project Page: https://yiwen-song.github.io/paper_orchestra/

1 Introduction

The rapid advancement of Large Language Models (LLMs) is transitioning AI from an assistive tool to an active participant in scientific discovery (Eger et al., 2025). While recent end-to-end autonomous frameworks (Lu et al., 2024; Yamada et al., 2025) establish the feasibility of automated research loops, realizing their full potential is hindered by a critical step: translating unstructured materials, such as raw ideas and experimental logs, into rigorous, submission-ready manuscripts.

Early attempts at automated academic writing relied on the parametric memory of LLMs, often leading to factual hallucinations. To mitigate this, recent frameworks employ retrieval-augmented generation (RAG) methods. Systems like AutoSurvey2 (Wu et al., 2025) and LiRA (Go et al., 2026) successfully automate long-form literature surveys by dividing cognitive tasks to improve structural coherence. However, these survey-specific frameworks lack the capacity to transform raw experimental data into a full-length research paper.

On the other hand, full-lifecycle autonomous research agents are tightly coupled to their experimental loops, preventing them from functioning as standalone writing tools capable of processing human-provided materials. Empirical evaluations show critical deficits in their literature synthesis (Beel et al., 2025; Tang et al., 2025b). Relying on simple keyword searches, these agents produce shallow reviews with insufficient citations. They also lack the capabilities to generate conceptual diagrams, restricting visuals to code-generated data plots. Furthermore, while comprehensive benchmarks like MLR-Bench (Chen et al., 2025a) have emerged to evaluate the full open-ended research lifecycle, evaluating the automated writing phase independently remains difficult due to the absence of a specialized, standardized benchmark.

To bridge these gaps, our core contributions are:

Method	E2E LaTeX Manuscript Gen.	Decoupled Standalone Writer	Robust to Unstructured Input	Targeted Lit Review Gen.	Conceptual Diagram Gen.
PaperRobot (Wang et al., 2019)	✗	✗	✗	✗	✗
data-to-paper (Technion-Kishony Lab, 2024)	✗	✓	✓	✗	✗
AI-Researcher (Tang et al., 2025a)	✓	✗	✗	✓	✗
CycleResearcher (Weng et al., 2025)	✓	✗	✗	✗	✗
AI Scientist-v2 (Yamada et al., 2025)	✓	✗	✓	✓	✗
PAPERORCHESTRA **(Ours)**	✓	✓	✓	✓	✓

Table 1: Comparison of PAPERORCHESTRA against existing AI writing systems.

- **PAPERORCHESTRA:** A standalone, multi-agent framework that autonomously authors LaTeX manuscripts from unconstrained pre-writing materials. It uses specialized agents to synthesize deep literature reviews, generate plots and conceptual diagrams, and iteratively refine the manuscript for better technical clarity.
- **`PaperWritingBench`:** The first standardized benchmark for AI research paper writing, which isolates the writing task by providing reverse-engineered raw materials (ideas and experimental logs) derived from 200 papers published in top-tier AI conferences.
- **Performance:** In side-by-side human evaluations, PAPERORCHESTRA significantly outperforms autonomous baselines, achieving absolute win rate margins (the difference of our win rate and the baseline's) of **50%–68%** in literature review synthesis and **14%–38%** in overall manuscript quality.

2 Related Work

2.1 AI Researcher Frameworks

Recent works have introduced end-to-end (E2E) frameworks designed to fully automate the scientific research lifecycle. The AI Scientist-v1 (Lu et al., 2024) introduced automated experimentation and drafting via code templates, while v2 (Yamada et al., 2025) achieved full autonomy using agentic tree-search. Concurrent works expand on these paradigms through distinct technical approaches: CycleResearcher (Weng et al., 2025) employs an iterative refinement loop with reinforcement learning from reviewer feedback, while Omni-Scientist (Shao et al., 2025) couples an agentic framework with structured knowledge and tool-based reasoning. Other systems, such as InternAgent (Zhang et al., 2025) and AI Co-Scientist (Gottweis et al., 2025), support human-in-the-loop collaboration in domain-specific applications. While effective at autonomous execution, the writing modules in these frameworks are rigidly coupled to their internal experimental loops. Consequently, they cannot directly function as standalone pipelines to synthesize unconstrained, human-provided pre-writing materials into submission-ready manuscripts.

2.2 Automated Writing and Literature Synthesis

Early automated writing systems, such as PaperRobot (Wang et al., 2019), generated incremental text sequences conditioned on entity relation graphs but lacked the capacity to synthesize complex, data-driven scientific narratives. More recent LLM-based writing assistants, including Prism (OpenAI, 2026), excel at stylistic refinement and local text editing, but typically rely on structured inputs or human guidance and do not natively support end-to-end manuscript generation from raw experimental logs. Furthermore, existing work (Gooding et al., 2025) explores writing as a testbed for evaluating open-ended, collaborative agents. However, their focus remains on broad writing evaluations rather than the rigid, structural constraints required for conference manuscript generation. Similarly, data-to-paper (Technion-Kishony Lab, 2024) focuses on translating structured analytical results into text with backward traceability. However, it primarily operates on structured, well-defined inputs and is less suited to flexibly drafting from early-stage, under-specified raw materials.

Beyond constructing a data-driven narrative from experimental results, synthesizing prior literature remains a distinct and critical challenge in automated authoring. Recent work

targets this problem via automated literature review generation. For instance, AutoSurvey2 (Wu et al., 2025) uses a multi-stage retrieval-augmented pipeline to generate comprehensive surveys, while SurveyGen-I (Chen et al., 2025b) focuses on iteratively refining topic outlines. Similarly, LiRA (Go et al., 2026) employs a RAG-based workflow to aggregate and summarize clusters of related papers. While highly effective at generating long-form surveys, these systems are not designed for writing targeted related work sections. Consequently, they often lack the contextual awareness needed to selectively contrast prior work and clearly motivate the specific research gap of a new method.

As summarized in Table 1, existing pipelines show key limitations when used as standalone manuscript generators. Frameworks like CycleResearcher (Weng et al., 2025) require a structured BibTeX reference list as input, which is rarely available at the start of writing, and fail on unstructured inputs outside this format. While AI-Researcher (Tang et al., 2025a) and AI Scientist-v2 (Yamada et al., 2025) accept unstructured topic descriptions, their writing modules rely on artifacts produced earlier in their internal pipelines and do not generate conceptual scientific diagrams. In contrast, PAPERORCHESTRA can process unconstrained pre-writing materials without relying on pre-existing plots or ground-truth references. By orchestrating specialized agents to autonomously synthesize targeted literature, generate comprehensive visuals, and craft coherent scientific narratives, PAPERORCHESTRA establishes a robust, end-to-end pipeline for AI research manuscript generation.

3 Task and Dataset

3.1 Task Formulation

We formulate the end-to-end AI research paper generation task as a function mapping unconstrained pre-writing materials to a complete submission package. Specifically, the framework operates on the following input components:

- **Idea Summary** ($\mathcal{I}$): A brief overview establishing the proposed methodology, core contributions, and theoretical foundation.
- **Experimental Log** ($\mathcal{E}$): A compilation of experimental results, covering raw data points, ablation studies, and performance metrics.
- **LaTeX Template** ($\mathcal{T}$): The template files provided by the target AI conference.
- **Conference Guidelines** ($\mathcal{G}$): The requirements mandated by the target AI conference.
- **Figures** ($\mathcal{F}$): An optional set of pre-existing visual assets (e.g., diagrams, plots). If no figures are provided ($\mathcal{F} = \emptyset$), the pipeline autonomously synthesizes all relevant visuals.

The goal is to produce a finalized submission package P. We define P as a tuple consisting of the source LaTeX file (P_{tex}) and the rendered PDF (P_{pdf}), generated via the framework W:

$$P = (P_{tex}, P_{pdf}) = W(\mathcal{I}, \mathcal{E}, \mathcal{T}, \mathcal{G}, \mathcal{F}) \tag{1}$$

3.2 Dataset Construction

To evaluate our pipeline, we introduce `PaperWritingBench` (App. D.1), a new dataset comprising 200 accepted papers from CVPR 2025 and ICLR 2025 (100 papers each). These venues ensure high academic standards while testing adaptability to distinct conference formats (double-column CVPR vs. single-column ICLR). To construct the evaluation tuple $(\mathcal{I}, \mathcal{E}, \mathcal{T}, \mathcal{G}, \mathcal{F})$ for each paper, we retrieve the official LaTeX templates ($\mathcal{T}$) and guidelines ($\mathcal{G}$) for each venue and execute the following pipeline:

Paper Acquisition and Extraction. We randomly sample papers from the OpenReview portal (ICLR) and the CVF Open Access repository (CVPR) to ensure topic diversity. We process the raw PDFs using MinerU (Wang et al., 2024) to generate structured, mathematically faithful Markdown, and PDFFigures 2.0 (Clark & Divvala, 2016) to extract visual entities ($\mathcal{F}$) and captions. Incomplete or misparsed samples are discarded to maintain corpus quality.

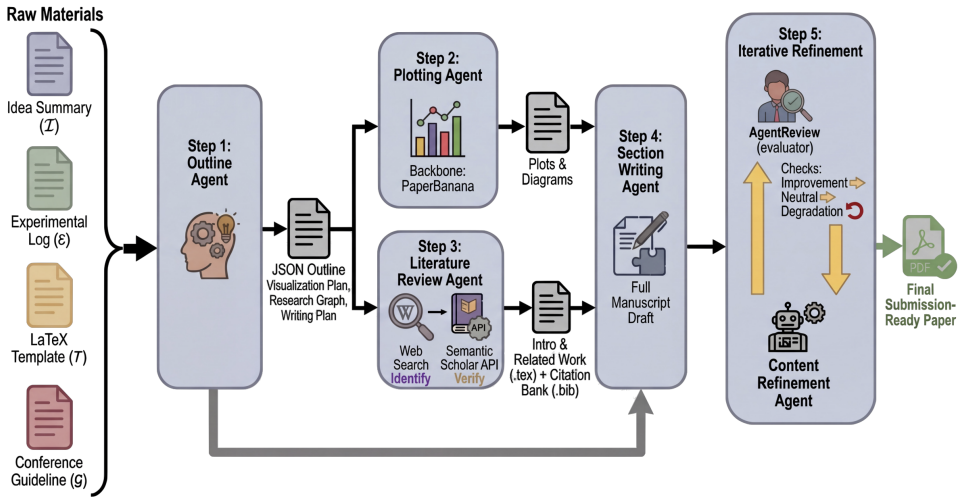


Figure 1: Overview of the PAPERORCHESTRA framework[1]. Specialized agents systematically parse unstructured inputs, synthesize plots and literature, compile a full draft, and iteratively refine the manuscript into a submission-ready PDF.

Synthesizing Raw Materials. Since pre-writing materials (e.g., lab notes) are unavailable, we prompt an LLM to reverse-engineer two core components from the extracted PDF text (App. D.2). To prevent information leakage, both components are fully anonymized (stripping authors and titles) and rendered strictly self-contained by removing all citations, URLs, and explicit figure or table references. We synthesize the following (App. D.4):

- **Idea Summary ($\mathcal{I}$):** Distills the core methodology while explicitly excluding experimental results. We generate a *Sparse* variant (summarizing only high-level ideas) and a *Dense* variant (retaining formal definitions and LaTeX equations) to simulate different degrees of user drafting effort.
- **Experimental Log ($\mathcal{E}$):** Extracts a record of experimental setup and empirical findings, including baselines, datasets, metrics, and tabular data. The LLM further de-contextualizes this data by converting visual insights into standalone factual observations, allowing us to test how well the agent can reconstruct the narrative purely from raw data.

4 PAPERORCHESTRA

Figure 1 illustrates PAPERORCHESTRA, a multi-agent framework (App. G.1) that autonomously transforms pre-writing materials, including Idea Summary ($\mathcal{I}$), Experimental Log ($\mathcal{E}$), LaTeX Template ($\mathcal{T}$), and Conference Guidelines ($\mathcal{G}$), into submission-ready manuscripts. The pipeline executes five steps, with Step 2 and Step 3 operating in parallel:

Step 1: Outline Generation. The *Outline Agent* synthesizes pre-writing materials into a JSON outline comprising: (a) a visualization plan dictating specific plot types, data sources, and aspect ratios; (b) a targeted literature search strategy establishing both macro-level context and micro-level methodology clusters to guide the Literature Review Agent in conducting searches and constructing a citation map, which is subsequently leveraged to draft the Introduction and Related Work sections; and (c) a section-level writing plan including high-level content bullets and a comprehensive list of citation hints for all core external dependencies, such as baselines, datasets, and metrics used in the paper.

Step 2: Plot Generation. The *Plotting Agent* executes the visualization plan to generate conceptual diagrams and statistical plots. We use *PaperBanana* (Zhu et al., 2026) as the

[1]This figure was generated using PaperBanana (Zhu et al., 2026).

default module, which employs a closed-loop refinement system where a VLM critic evaluates rendered images against design objectives, iteratively revising text descriptions and regenerating images to resolve visual artifacts, and synthesizing context-aware captions.

Step 3: Literature Review. Executing the search strategy outlined in Step 1, the *Literature Review Agent* drives a concurrent, hybrid discovery pipeline. It employs an LLM equipped with web search to identify candidate papers, then uses the Semantic Scholar API (Kinney et al., 2023) to authenticate their existence (App. E.4). Upon a successful match, the agent fetches the abstract and metadata while enforcing temporal cutoffs (App. E.1); candidates are discarded if they exceed the cutoff date or lack a verified mapping. Following deduplication via Semantic Scholar IDs, the agent compiles a citation registry and auto-generates a BibTeX (`.bib`) file. Finally, it uses the verified citations to draft the *Introduction* and *Related Work* sections.

Step 4: Section Writing. The *Section Writing Agent* drafts the remaining core sections using the outputs from previous stages. Building upon the partially filled LaTeX file, it extracts numeric values from the experimental log to construct tables. Guided by the section-level outline and the established citation bank, the agent authors the abstract, methodology, experiments, and conclusion sections, seamlessly integrating the generated figures to produce a complete LaTeX manuscript.

Step 5: Iterative Content Refinement. The *Content Refinement Agent* iteratively optimizes the manuscript using simulated peer-review feedback. Here we use *AgentReview* (Jin et al., 2024) as the default system in this module. After modifying the LaTeX source to address weaknesses, revisions are accepted if the overall score increases, or if it ties while net sub-axis gains are non-negative. The agent immediately reverts to the previous version and halts upon any overall score decrease, negative tie-breaker, or reaching the iteration limit. The resulting LaTeX document and compiled PDF represent the final output of PAPERORCHESTRA.

5 Experiments

5.1 Baselines

We benchmark PAPERORCHESTRA against two pipelines (App. E.2): (1) a **Single Agent** baseline that processes all raw materials ($\mathcal{I}, \mathcal{E}, \mathcal{T}, \mathcal{G}, \mathcal{F}$) and executes end-to-end drafting and bibliography construction in a single LLM call, validating the necessity of a multi-agent system; and (2) **AI Scientist-v2** (Yamada et al., 2025), a state-of-the-art system featuring multi-round citation gathering, VLM-guided plot refinement, and iterative self-reflection. We also evaluate a compute-matched AgentLaboratory (Schmidgall et al., 2025) and a general-purpose Google Antigravity (Google, 2026) baseline in App. E.3.

To ensure fair comparison against baselines lacking diagram generation, we standardize visual inputs by restricting allowable figures to those extracted from ground-truth papers ($\mathcal{F} = \mathcal{F}_{GT}$). Furthermore, we adopt the *sparse idea* as our standard setting to simulate realistic scenarios where researchers provide unstructured, preliminary notes rather than mature formulations. To prevent pre-training data memorization, all evaluated pipelines employ a universal anti-leakage prompt (App. E.5).

5.2 Autoraters

Citation F1. To quantitatively assess citation coverage, we compare the generated reference lists against the ground-truth (GT) paper. Using an LLM provided with the paper context (see App. G.3), we partition the GT references into two categories: (1) **P0 (Must-Cite)** comprises core citations strictly necessary for contextualizing the work, including direct experimental baselines, utilized datasets, metrics, and foundational methodologies upon which the paper directly builds; (2) **P1 (Good-to-Cite)** includes all remaining GT citations, providing valuable but non-essential background context such as orthogonal work or supplementary information. To ensure accurate matching, we extract reference lists from

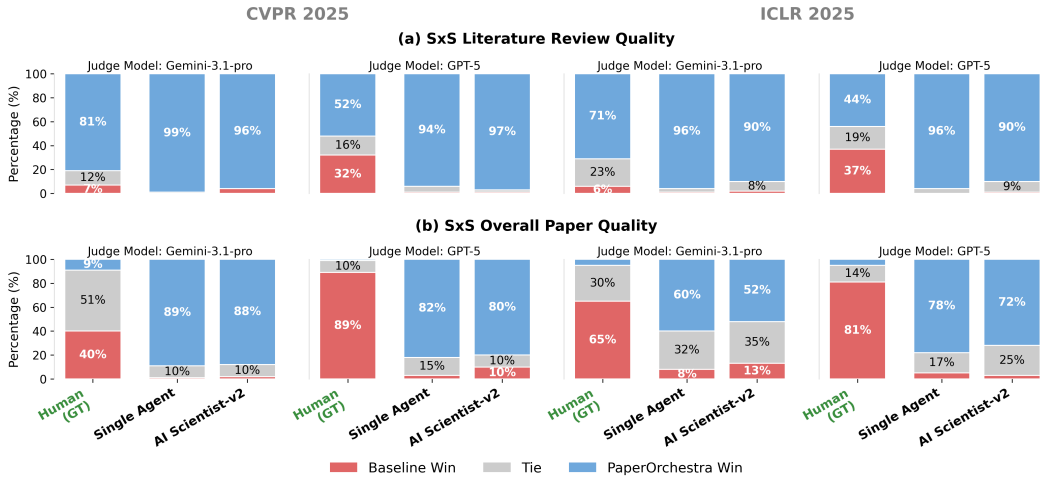


Figure 2: Automated SxS evaluation of PAPERORCHESTRA against baselines on the CVPR and ICLR 2025 datasets under the sparse idea setting. The **Human (GT)** baseline serves as an upper-bound reference. For both **(a)** literature review and **(b)** overall paper quality, PAPERORCHESTRA significantly outperforms the fully autonomous baselines (Single Agent and AI Scientist-v2) across both judge models.

both the GT and generated papers and resolve them to unique entity IDs using the Semantic Scholar API. We then compute the Recall and F1 scores independently for the P0 set, the P1 set, and the combined overall reference set.

Literature Review Quality. To qualitatively evaluate the generated *Introduction* and *Related Work* sections, we employ an LLM evaluator to score the generated text from 0–100 across six axes: Coverage & Completeness, Relevance & Focus, Critical Analysis & Synthesis, Positioning & Novelty, Organization, and Citation Rigor. To mitigate standard LLM score inflation, the evaluation is grounded against venue-specific average citation counts. The system enforces strict score caps and penalties for purely descriptive summaries, unsupported novelty claims, or citation padding, outputting an overall quality score for each paper.

Overall Quality. To evaluate the holistic technical quality of the generated papers, we employ two AI-based peer review frameworks simulating expert assessment: (1) the **AI Scientist-v2 Reviewer** (Yamada et al., 2025), an automated module for structured manuscript evaluation; and (2) **ScholarPeer** (Goyal et al., 2026), a search-enabled multi-agent system mimicking expert workflows via iterative retrieval and evidence checking. Both systems yield multi-axis scores, an overall rating, and a simulated acceptance decision.

Side-by-Side (SxS) Comparison. To directly compare manuscripts generated from the same research idea across different pipelines, we implement two automated SxS LLM evaluators. **(1) SxS Literature Review Quality** extracts the *Introduction* and *Related Work* to assess problem framing, prior work coverage, organization and synthesis, contribution positioning, and readability. **(2) SxS Paper Quality** holistically compares the full manuscript (including visual layouts) across six axes: scientific depth, technical execution, logical flow, writing clarity, presentation of evidence, and academic style. To mitigate LLM positional bias, we evaluate each pair in both orderings. The final aggregated outcome is recorded as a *win* (two wins, or one win and one tie), a *tie* (one win and one loss, or two ties), or a *loss*.

5.3 Results

Side-by-Side Comparison. As shown in Figure 2, PAPERORCHESTRA consistently outperforms both the Single Agent baseline and AI Scientist-v2 in SxS evaluations. In *Literature Review Quality*, our framework dominates the autonomous baselines, achieving absolute win

Venue	Pipeline	Originality (1-4)	Quality (1-4)	Clarity (1-4)	Significance (1-4)	Soundness (1-4)	Presentation (1-4)	Contribution (1-4)	Overall (1-10)	Acceptance Rate (%)
AI Scientist-v2 Reviewer (Yamada et al., 2025)										
CVPR 2025	Original (GT)	2.96	**2.81**	**3.44**	**3.06**	**2.74**	**3.26**	**2.91**	**5.95**	**71.00**
	Single Agent	**3.01**	2.27	2.78	3.03	2.26	2.79	2.74	4.60	33.00
	AI Scientist-v2	2.93	2.11	2.80	2.85	2.15	2.82	2.61	4.22	22.00
	PaperOrchestra (PlotOff)	2.95	2.43	3.40	2.97	2.42	3.22	2.74	5.12	48.00
ICLR 2025	Original (GT)	**2.99**	**2.79**	**3.41**	**3.03**	**2.76**	**3.30**	**2.94**	**5.81**	**63.00**
	Single Agent	2.70	1.67	2.20	2.44	1.76	2.19	2.09	3.22	4.00
	AI Scientist-v2	2.58	1.76	2.27	2.46	1.82	2.29	2.16	3.42	11.00
	PaperOrchestra (PlotOff)	2.68	2.10	2.82	2.57	2.12	2.79	2.37	4.10	22.00
ScholarPeer Reviewer (Goyal et al., 2026)										
CVPR 2025	Original (GT)	**3.22**	**3.02**	3.54	**3.63**	2.87	3.39	**3.42**	**7.07**	**86.00**
	Single Agent	**3.22**	2.79	3.46	3.48	2.72	3.34	3.29	6.35	71.00
	AI Scientist-v2	3.11	2.71	3.44	3.39	2.64	3.37	3.22	6.30	70.00
	PaperOrchestra (PlotOff)	3.13	2.95	**3.75**	3.56	**2.91**	**3.64**	3.34	6.93	84.00
ICLR 2025	Original (GT)	**3.33**	**3.14**	**3.76**	**3.74**	**3.05**	**3.59**	**3.57**	**7.48**	**94.00**
	Single Agent	3.26	2.73	3.27	3.48	2.71	3.20	3.25	6.35	72.00
	AI Scientist-v2	3.23	2.62	3.28	3.41	2.60	3.23	3.11	6.04	64.00
	PaperOrchestra (PlotOff)	3.35	2.96	3.65	3.59	2.94	3.48	3.37	7.03	81.00

Table 2: Holistic technical quality evaluation via the AI Scientist-v2 (Yamada et al., 2025) and ScholarPeer (Goyal et al., 2026) frameworks (Sparse Idea Setting). Higher scores indicate better manuscript quality and increased simulated acceptance rates. **Bold** indicates best scores; while green highlights the highest score achieved by an AI pipeline.

Venue	Pipeline	GT Avg # Cites	Avg # Cites	Gemini-3.1-Pro			GPT-5		
				Overall F1	P0 Recall	P1 Recall	Overall F1	P0 Recall	P1 Recall
CVPR 2025	Single Agent	58.52	11.46	28.03	60.84	2.77	28.16	63.04	3.27
	AI Scientist-v2		14.18	17.15	36.12	3.26	17.26	37.46	3.30
	PaperOrchestra (PlotOff)		47.98	**29.65**	**63.58**	**15.85**	**29.92**	**65.17**	**16.43**
ICLR 2025	Single Agent	59.18	9.75	23.16	48.41	3.09	23.29	50.75	3.23
	AI Scientist-v2		13.71	15.80	33.66	3.07	15.94	34.73	3.36
	PaperOrchestra (PlotOff)		45.73	**27.76**	**54.48**	**16.43**	**27.97**	**55.70**	**17.11**

Table 3: Citation F1 evaluation (Sparse Idea Setting). P0 and P1 denote must-cite and good-to-cite references, respectively. F1 and Recall are percentages.

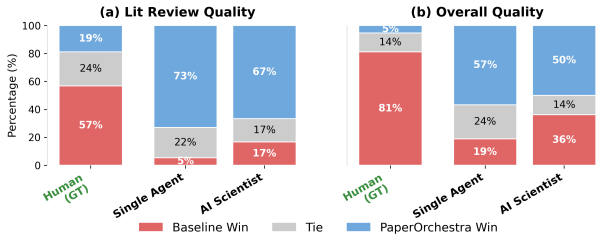
margins of **88%–99%**. For *Overall Paper Quality*, although Human (GT) remains the upper bound, PaperOrchestra substantially surpasses all tested AI competitors. It strongly outperforms AI Scientist-v2 and the Single Agent by margins of **39%–86%** and **52%–88%**, respectively, across all settings, confirming that our multi-agent architecture significantly enhances overall manuscript quality (cost analysis in App. C).

Technical Quality. As detailed in Table 2, we evaluate holistic manuscript quality using the AI Scientist-v2 and ScholarPeer frameworks. Under ScholarPeer, PaperOrchestra achieves high simulated acceptance rates of **84%** (CVPR) and **81%** (ICLR), closely tracking Human (GT) baselines (**86%** and **94%**, respectively). It significantly outperforms existing AI pipelines, delivering absolute acceptance gains of **13%** (CVPR) and **9%** (ICLR) over the strongest autonomous baseline. Beyond acceptance rates, PaperOrchestra dominates across critical sub-axes (notably *Clarity*, *Presentation*, and *Soundness*), demonstrating that our multi-agent system yields structurally coherent and technically sound manuscripts.

Citation Coverage. Table 3 demonstrates PaperOrchestra successfully mitigates under-citation issues. While autonomous baselines achieve competitive Overall F1 scores, this is a mathematical artifact of their extremely low citation counts (averaging 9–14). They focus on fetching obvious P0 papers, which inflates their F1 scores but leaves P1 Recall near zero. In contrast, our framework generates **45.73–47.98** citations, closely mirroring Human (GT) writeups ($\sim$59). Alongside improving P0 Recall by absolute margins of **2.13%–6.07%**, PaperOrchestra significantly increases P1 Recall by **12.59%–13.75%** over the strongest baselines. This proves our method actively explores the broader academic landscape rather than relying on shallow keyword matching.

Venue	Pipeline	Citation Practices	Coverage & Completeness	Critical Analysis & Synthesis	Organization & Writing	Positioning & Novelty	Relevance & Focus	Overall Score
Judge Model: Gemini-3.1-Pro								
CVPR 2025	Human (GT)	69.88	**71.55**	65.80	74.31	75.55	78.25	71.08
	Single Agent	36.83	30.10	52.93	77.03	66.45	71.40	43.04
	AI Scientist-v2	39.24	34.25	54.83	77.21	65.16	69.12	45.43
	PaperOrchestra (PlotOff)	**75.36**	70.15	**79.01**	**83.74**	**82.68**	**81.15**	**78.30**
ICLR 2025	Human (GT)	**71.66**	**68.30**	74.60	80.54	**82.22**	**80.67**	75.18
	Single Agent	33.03	26.35	49.80	75.38	61.14	69.70	38.91
	AI Scientist-v2	37.60	32.27	51.84	75.74	62.96	67.20	42.98
	PaperOrchestra (PlotOff)	71.15	65.71	**77.34**	**82.02**	81.94	80.08	**76.23**
Judge Model: GPT-5								
CVPR 2025	Human (GT)	63.47	63.86	57.36	76.16	**61.34**	**73.23**	53.66
	Single Agent	49.69	47.01	54.49	76.55	56.62	70.60	44.33
	AI Scientist-v2	49.18	47.59	53.71	75.71	55.80	69.00	43.57
	PaperOrchestra (PlotOff)	**63.88**	**64.13**	**60.04**	**77.46**	59.18	72.65	**53.99**
ICLR 2025	Human (GT)	61.61	60.91	58.75	77.19	**62.91**	73.11	**55.53**
	Single Agent	49.34	45.97	54.10	77.79	56.47	71.52	44.30
	AI Scientist-v2	49.55	48.26	54.18	76.62	56.08	69.86	44.29
	PaperOrchestra (PlotOff)	**62.54**	**62.12**	**60.58**	**78.70**	60.21	**73.58**	54.15

Table 4: Literature review quality assessed by LLM-as-a-Judge (0–100 scale, Sparse Idea Setting).



(a) Human SxS Evaluation Results

Metric	Judge(s)	Pearson	Spearman
Human vs. Autorater Correlation			
Lit Review Quality	Gemini-3.1-Pro	0.0868	0.0715
	GPT-5	**0.2764**	**0.2300**
Overall Quality	Gemini-3.1-Pro	0.5672	0.5727
	GPT-5	**0.6458**	**0.6355**
Inter-Autorater Correlation (Gemini-3.1-Pro vs. GPT-5)			
Lit Review Quality	–	0.5697	0.5615
Overall Quality	–	**0.7395**	**0.7454**

(b) Autorater Validation Metrics

Figure 3: Human & Autorater Evaluation (Sparse Idea Setting). **(a)** Human preferences for PaperOrchestra vs. baselines. **(b)** GPT-5 strongly correlates with human scores on overall quality, alongside high inter-autorater consistency between distinct LLM judges.

Literature Review Quality. Citation F1 alone cannot fully evaluate literature synthesis; a review must also identify gaps and motivate the proposed method. From Table 4, PaperOrchestra achieves absolute Overall Score gains of **32.87%–33.25%** (Gemini-3.1-Pro) and **9.66%–9.85%** (GPT-5) over the strongest AI baseline, remaining highly comparable to Human baselines. Dominating in *Citation Practices* and *Critical Analysis*, our framework synthesizes analytical, well-grounded narratives rather than generic LLM summaries.

5.4 Human Evaluation

To validate automated SxS metrics and assess human-perceived manuscript quality, we recruited 11 AI researchers to evaluate 40 randomly sampled papers (20 per venue) from `PaperWritingBench`. PaperOrchestra was compared against all baselines (Human GT, Single Agent, and AI Scientist-v2), yielding 180 paired evaluations (App. E.6).

Human preferences correlate strongly with our GPT-5 evaluator for Overall Quality (Pearson $r = 0.6458$, Spearman $\rho = 0.6355$; Figure 3b). Literature review correlation is lower due to inherent LLM self-bias. Manual inspection reveals LLMs tend to act as structural graders, rewarding rigid formatting such as explicit "Problem-Gap-Solution" paragraphs or bulleted thematic groups. In contrast, human experts prioritize dense, pragmatic factuality and a nuanced, narrative-driven flow, which the autorater often mistakenly penalizes for lacking explicit formatting cues. Nevertheless, relative rankings remain strictly consistent. The stability of our automated SxS metrics is further validated by high inter-autorater agreement across different LLM judges (Figure 3b). Human SxS evaluations (Figure 3a) confirm PaperOrchestra strongly outperforms AI baselines by absolute margins of **50%–68%**

Venue	Metric	Gemini-3.1-Pro			GPT-5		
		Sparse Win	Tie	Dense Win	Sparse Win	Tie	Dense Win
CVPR 2025	Paper Quality	18%	26%	**56%**	24%	32%	**44%**
	Lit Review	37%	24%	**39%**	**40%**	22%	38%
ICLR 2025	Paper Quality	24%	31%	**45%**	20%	37%	**43%**
	Lit Review	33%	33%	**34%**	32%	**40%**	28%

Table 5: Automated SxS evaluation of Sparse vs. Dense idea settings.

Venue	Gemini-3.1-Pro			GPT-5		
	PlotOff Win	Tie	PlotOn Win	PlotOff Win	Tie	PlotOn Win
CVPR 2025	**34%**	32%	34%	**49%**	19%	32%
ICLR 2025	**35%**	37%	28%	**36%**	38%	26%

Table 6: Automated SxS Paper Quality evaluating the Plotting Agent. Despite lacking access to data represented by GT visuals, *PlotOn* achieves highly competitive performance against *PlotOff* that use human-authored figures.

(Literature Review) and **14%–38%** (Overall Quality). It also achieves a highly competitive **43%** tie/win rate against human GT in literature synthesis.

5.5 Ablation Studies

Sparse vs. Dense Idea Table 5 evaluates the impact of input density. For *Overall Paper Quality*, the Dense idea setting dominates win rates (**43%–56%** vs. **18%–24%**) by facilitating more rigorous methodology generation. Meanwhile, *Literature Review Quality* exhibits near-parity, with the Sparse setting securing **32%–40%** against Dense's **28%–39%**. This demonstrates the exceptional robustness of PAPERORCHESTRA: our *Literature Review Agent* autonomously executes targeted searches and identifies research gaps without relying heavily on the dense input drafted by human. Additional qualitative comparisons are provided in App. F.2.

Autonomous Visual Generation (PlotOff vs. PlotOn) Table 6 evaluates our *Plotting Agent* by comparing papers using human-authored GT figures (*PlotOff*) vs. autonomously generated visuals (*PlotOn*). Although *PlotOff* benefits from an inherent information advantage (GT figures often embed supplementary results absent from raw input logs), *PlotOn* remains highly competitive, securing ties or wins in **51%–66%** of SxS matchups across all four setups. As shown in App. F.1, PAPERORCHESTRA synthesizes coherent visuals from scratch, effectively augmenting the manuscript and reinforcing the scientific narrative.

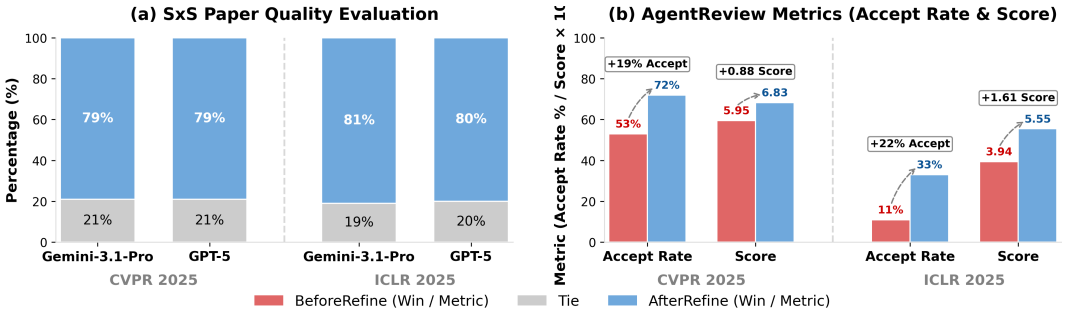


Figure 4: **Impact of the Content Refinement Agent.** (a) In automated SxS paper quality evaluation, refined manuscripts dominate. (b) This improvement is reflected through absolute gains in acceptance rates simulated by AgentReview (Jin et al., 2024).

Importance of the Content Refinement Agent Figure 4 isolates the impact of our iterative evaluation-feedback loop. In automated SxS comparisons (4(a)), refined manuscripts (*AfterRefine*) dominate unrefined drafts (*BeforeRefine*) with **79%–81%** win rates and **0%** losses. AgentReview metrics (4(b)) reflect this trend, showing substantial increases in simulated acceptance rates (**+19%** CVPR, **+22%** ICLR) and overall scores (**+0.88** and **+1.61**). Driven by targeted clarity and presentation corrections, these results prove our content refinement loop is critical for elevating raw drafts into rigorous, submission-ready manuscripts.

Pipeline Modularity & Refinement Oracles To verify that PAPERORCHESTRA does not depend on a specific feedback oracle, we replaced our default AgentReview module with the AI Scientist-v2 Reviewer. As shown in Table 7, both oracles generate highly competitive papers with comparable simulated acceptance rates (84% vs. 87% under ScholarPeer evaluation), demonstrating the plug-and-play nature of our refinement mechanism.

Content Refiner	Originality (1-4)	Quality (1-4)	Clarity (1-4)	Significance (1-4)	Soundness (1-4)	Presentation (1-4)	Contribution (1-4)	Overall (1-10)	Acceptance Rate (%)
AgentReview (Default)	3.13	2.95	3.75	3.56	2.91	3.64	3.34	6.93	84.00
AI Scientist-v2 Reviewer	3.24	3.02	3.64	3.70	2.96	3.56	3.51	7.34	87.00

Table 7: ScholarPeer evaluation of PAPERORCHESTRA using different refinement oracles on CVPR 2025.

Citation Verification Sensitivity We analyzed the sensitivity of PAPERORCHESTRA's citation performance by varying the Levenshtein ratio threshold for reference matching. As detailed in Table 8, both objective citation F1 and subjective Literature Review Quality remained highly stable across a broad parameter space (50%–90%). We selected 70% as the default because it empirically yielded the highest Literature Review Quality Overall Score (78.30%) while maintaining a near-optimal Overall F1 score (29.65%).

Threshold (%)	Overall F1 (%)	Lit. Review Quality Score
50	28.21	78.02
60	29.22	78.17
70 (Default)	29.65	**78.30**
80	**29.71**	77.54
90	28.67	76.27

Table 8: Citation Verification Threshold Ablation Results.

6 Conclusion

In this work, we introduced PAPERORCHESTRA and `PaperWritingBench` to transform unstructured, preliminary AI research materials into submission-ready manuscripts. Experiments demonstrate that our multi-agent framework can generate high-quality research papers with competitive runtime (App. C) while synthesizing deep, context-aware literature reviews. Future research can focus on expanding the framework's capacity to ingest richer research artifacts and transitioning these autonomous agents into interactive, dynamic writing environments to enable seamless human-AI scientific collaboration (more limitations discussed in App. A).

Ethics Statement

We position our system as an advanced assistive tool designed to accelerate the drafting process of AI research papers, rather than an independent entity capable of claiming authorship. Human researchers must retain full accountability for the factual accuracy, originality, and validity of the claims presented in any generated manuscript. While PAPERORCHESTRA incorporates robust programmatic safeguards (such as API-grounded citation validation) to minimize hallucinations and ensure academic rigor, users are responsible for verifying the outputs to prevent the propagation of LLM-derived biases or misinformation.

References

Anthropic. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, Feb 2026. Accessed August 7, 2026.

Mohammadhossein Bahari, Saeed Saadatnejad, Amirhossein Askari Farsangi, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Certified human trajectory prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12301–12311, 2025.

Joeran Beel, Min-Yen Kan, and Moritz Baumgart. Evaluating sakana's AI scientist: Bold claims, mixed results, and a promising future? *ACM SIGIR Forum*, 59(1):1–20, 2025. doi: 10.1145/3769733.3769747. URL https://doi.org/10.1145/3769733.3769747.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. The neurips 2021 consistency experiment. NeurIPS Blog, December 2021. URL https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. MLR-Bench: Evaluating AI agents on open-ended machine learning research. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-3945. Datasets and Benchmarks Track.

Jing Chen, Zhiheng Yang, Yixian Shen, Jie Liu, Adam Belloum, Paola Grosso, and Chrysa Papagianni. Surveygen-i: Consistent scientific survey generation with evolving plans and memory-guided writing. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 3687–3714, 2025b.

Christopher Clark and Santosh Divvala. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pp. 143–152. Association for Computing Machinery, 2016. ISBN 9781450342292. doi: 10.1145/2910896.2910904. URL https://doi.org/10.1145/2910896.2910904.

Darko Drakulić, Sofia Michel, and Jean-Marc Andreoli. Goal: A generalist combinatorial optimization agent learner. In *International Conference on Learning Representations*, volume 2025, pp. 52465–52488, 2025.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, et al. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *arXiv preprint arXiv:2502.05151*, 2025.

Gregory Hok Tjoan Go, Khang Ly, Anders Søgaard, Seyed Amin Tabatabaei, Maarten de Rijke, and Xinyi Chen. LiRA: A multi-agent framework for reliable and readable literature review generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 40456–40464, 2026. doi: 10.1609/aaai.v40i47.41489.

Sian Gooding, Lucia Lopez-Rivilla, and Edward Grefenstette. Writing as a testbed for open ended agents. *arXiv preprint arXiv:2503.19711*, 2025.

Google. Google antigravity. https://antigravity.google/, 2026. Accessed August 7, 2026.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Palash Goyal, Mihir Parmar, Yiwen Song, Hamid Palangi, Tomas Pfister, and Jinsung Yoon. Scholarpeer: A context-aware multi-agent framework for automated peer review. *arXiv preprint arXiv:2601.22638*, 2026.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1208–1226, 2024.

K-Dense-AI. Claude Scientific Writer. GitHub repository, 2026. URL https://github.com/K-Dense-AI/claude-scientific-writer. Accessed August 7, 2026.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

OpenAI. Introducing prism. https://openai.com/index/introducing-prism/, 2026. Accessed March 16, 2026.

Abduljalil Radman and Jorma Laaksonen. Tsam: Temporal sam augmented with multimodal prompts for referring audio-visual segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23947–23956, 2025. doi: 10.1109/CVPR52734.2025.02230.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, November 2025. URL https://aclanthology.org/2025.findings-emnlp.320/.

Chenyang Shao, Dehao Huang, Yu Li, Keyu Zhao, Weiquan Lin, Yining Zhang, Qingbin Zeng, Zhiyu Chen, Tianxing Li, Yifei Huang, et al. Omniscientist: Toward a co-evolving ecosystem of human and ai scientists. *arXiv preprint arXiv:2511.16931*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. AI-Researcher: Autonomous scientific innovation. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-0320.

Xuemei Tang, Xufeng Duan, and Zhenguang Cai. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1602–1617, 2025b.

Technion-Kishony Lab. data-to-paper, 2024. URL https://github.com/Technion-Kishony-lab/data-to-paper. Accessed March 16, 2026.

The Gemini Team. Gemini 3.1 pro. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026. Large language model accessed via API.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. Mineru: An open-source solution for precise document content extraction, 2024. URL https://arxiv.org/abs/2409.18839.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. Paperrobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1980–1991, 2019.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. CycleResearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*, 2025.

Siyi Wu, Chiaxin Liang, Ziqian Bi, Leyi Zhao, Tianyang Wang, Junhao Song, Yichao Zhang, Keyu Chen, Benji Peng, and Xinyuan Song. Autosurvey2: Empowering researchers with next level automated literature surveys. *arXiv preprint arXiv:2510.26012*, 2025.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Runmin Ma, Yusong Hu, Zhiyin Yu, Xiaohan He, Songtao Huang, et al. Internagent: When agent becomes the scientist–building closed-loop system from hypothesis to verification. *arXiv preprint arXiv:2505.16938*, 2025.

Dawei Zhu, Rui Meng, Yale Song, Xiyu Wei, Sujian Li, Tomas Pfister, and Jinsung Yoon. Paperbanana: Automating academic illustration for ai scientists. *arXiv preprint arXiv:2601.23265*, 2026.

Index of Appendices

- **App. A**: Limitations and Future Work
- **App. B**: Qualitative Analysis and Failure Modes
- **App. C**: Computational Cost
- **App. D**: Dataset Details
  - **App. D.1**: Data Distribution
  - **App. D.2**: Raw Material Extraction
  - **App. D.3**: Validity of Synthesized Input Materials
  - **App. D.4**: Raw Material Example
- **App. E**: Experiment Details
  - **App. E.1**: Models and Configuration
  - **App. E.2**: Baseline Selection
  - **App. E.3**: Additional Baselines
  - **App. E.4**: Citation Verification
  - **App. E.5**: Information Leakage Prevention
  - **App. E.6**: Human Evaluation
- **App. F**: Paper Visualization
  - **App. F.1**: Side-by-Side (SxS) Paper Visualizations
  - **App. F.2**: Sparse vs. Dense Idea Settings
- **App. G**: Prompts
  - **App. G.1**: Paper Writing
  - **App. G.2**: Raw Material Generation
  - **App. G.3**: Autoraters

A Limitations and Future Work

While PAPERORCHESTRA advances the capabilities of autonomous manuscript generation, several areas remain for future exploration. First, relying on external frameworks like PaperBanana (Zhu et al., 2026) for visual generation limits our direct control over figure hallucinations. Although mitigating these errors falls outside the primary scope of our pipeline, future systems could benefit from integrating targeted VLMs and dedicated human evaluations to systematically verify that generated visual content is factually sound and optimally placed within the LaTeX layout.

Second, while PAPERORCHESTRA significantly outperforms existing AI baselines, a quality gap remains when compared directly against human-written papers. We view this gap as a testament to the rigor of `PaperWritingBench`. To bridge it, transitioning our current framework toward an interactive, human-in-the-loop (HITL) system would enable researchers to iteratively steer drafts via natural language critiques. This evolution would solidify the framework's intended role as an advanced assistive tool, combining the efficiency of our autonomous baseline with the gold-standard rigor of human experts.

Third, `PaperWritingBench` focuses on CVPR and ICLR, representing the AI and machine learning domains. Scientific disciplines such as biology, physics, or the social sciences follow distinct writing conventions. Validating cross-domain transferability, such as adapting PAPERORCHESTRA to other fields by swapping paper retrieval endpoints and updating the auto-reviewing modules with discipline-specific guidelines and knowledge bases, remains an important direction for future work.

Finally, evaluating foundation models on `PaperWritingBench` carries an inherent risk of pretraining data contamination. To mitigate this risk, we rigorously de-contextualized and anonymized all input materials. Additionally, since all evaluated baselines share the same underlying language models, any potential knowledge leakage applies uniformly,

ensuring a strictly fair relative comparison. To fully isolate generation from memorization, future benchmarks could leverage unpublished research or autonomously generated raw materials.

B Qualitative Analysis and Failure Modes

To provide qualitative insights into PAPERORCHESTRA's operation and agent interactions, we examine its generation behavior across multiple generated manuscripts from `PaperWritingBench`. This analysis highlights the structural advantages of our multi-agent architecture, as well as its remaining failure modes.

B.1 Key Qualitative Successes

- **Accurate Claim Scoping:** The Outline Agent strictly maps downstream writing tasks to the provided empirical logs. This prevents the writing agents from "over-claiming" generalizations or asserting performance gains not present in the data, successfully mitigating the tendency to hallucinate results observed in the Single Agent baseline.
- **Handling Unactionable Feedback:** When the simulated reviewers request additional experiments, the Content Refinement Agent is explicitly instructed to focus on text clarification, structural revisions, and baseline presentation. By ignoring requests for post-facto experiments, it prevents hallucinating fake results or promising unfeasible future experiments.
- **Targeted Literature Reviews:** By isolating literature discovery from outline construction, the Outline Agent defines distinct search boundaries (macro-level context for the Introduction and micro-level methodological clusters for Related Work). This targeted search ensures every detected paper serves a precise narrative purpose.

B.2 Remaining Failure Modes

- **Citation Gaps:** The system occasionally misses highly foundational, ubiquitous literature (e.g., standard optimizers or basic backbones), assuming they are too well-known to require a dedicated query. Furthermore, since search results are verified exclusively via the Semantic Scholar database, technical reports or online documentation missing from the API cannot be resolved or cited.
- **Refinement Regressions:** In later editing rounds, aggressive polishing by the Content Refinement Agent can occasionally dilute technical specifics or introduce minor formatting and compile-time LaTeX errors if it drifts too far from the original outline. We restrict these regressions by capping content refinement at a maximum of three rounds.

C Computational Cost

Table 9 compares the computational cost and single-paper latency of PAPERORCHESTRA against the Single Agent and AI Scientist-v2 baselines. Although the Single Agent executes rapidly, it fails to produce the rigorous citations and data-grounded visuals generated by our system. Despite requiring more LLM calls ($\sim$60–70) than AI Scientist-v2 ($\sim$40–45), PAPERORCHESTRA maintains a highly competitive mean processing time of 39.6 minutes (compared to 35.1 minutes for AI Scientist-v2). To calculate these mean latency values, we evaluated all methods on a random subset of 10 papers (5 from CVPR and 5 from ICLR) using a single worker. For each method, we removed the highest and lowest execution times to mitigate the impact of API network anomalies before computing the final average.

The efficiency of PAPERORCHESTRA is achieved by strategically decoupling the paper discovery and verification pipeline. To optimize throughput, we execute retrieval in two distinct stages: (1) **Parallel Candidate Discovery**, which leverages 10 concurrent workers for search-grounded LLM calls to rapidly pool candidate papers; and (2) **Sequential Citation Verification**, which safely processes the pooled candidates through Semantic Scholar at the maximum allowable rate (1 query per second). This architecture successfully combines the

Method	#LLM Calls	Latency (Mean)	System Description
Single Agent	1	3.3 mins	Single monolithic prompt and generation.
AI Scientist-v2	∼40–45	35.1 mins	Multi-step pipeline (20-round citation scan, drafting, 3× reflections, and vision critique).
PAPERORCHESTRA	∼50–70	39.6 mins	Multi-agent orchestration (outline generation, parallel literature search, sequential citation verification, writing, plotting with critique, and a 3× content refinement loop).

Table 9: Comparison of computational cost and single-paper latency. This table outlines the estimated number of LLM API calls, mean execution time, and core architectural components for each pipeline.

high-concurrency tolerance of the LLM API with the strict throughput limits of the Semantic Scholar API to prevent quota-induced latency.

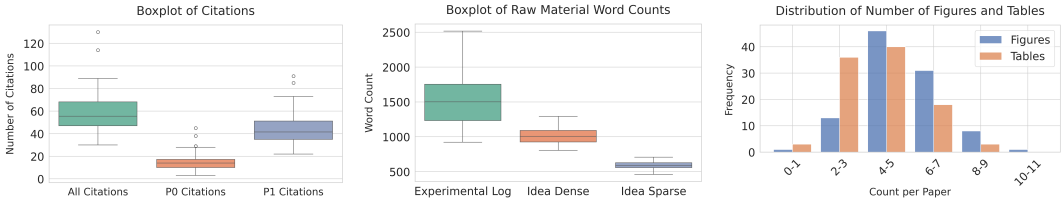
The LLM calls are distributed across our multi-agent system as follows:

- **Outline Agent** (1 call): Initial manuscript structuring and planning.
- **Hybrid Literature Agent** (∼20–30 calls): Execution of the decoupled paper discovery and citation verification pipeline described above.
- **Plotting Agent** (∼20–30 calls): Few-shot retrieval, visual planning, image generation, VLM-guided critique and redraw cycles, and context-aware captioning.
- **Section Writing Agent** (1 call): A single, comprehensive multimodal call to draft and compile the complete LaTeX manuscript.
- **Content Refinement Agent** (∼5–7 calls): Score-driven, iterative reflection and manuscript revision.

Ultimately, this computational investment is well justified. PAPERORCHESTRA delivers substantial improvements in overall manuscript quality and citation depth, while adding only minimal latency compared to the AI Scientist-v2 baseline.

D Dataset Details

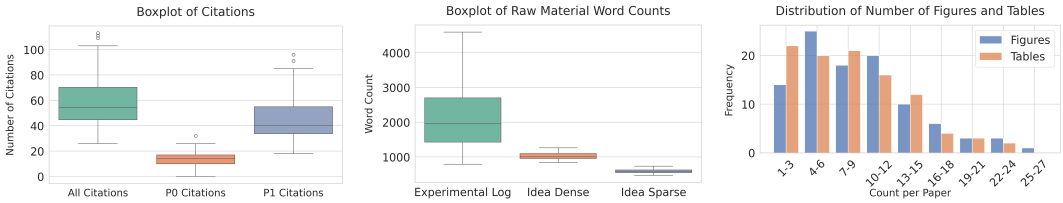
D.1 Data Distribution



(a) Citation Count Distribution (b) Material Length Distribution (c) Tables/Figures Distribution

Figure 5: CVPR 2025 Dataset Statistics

In this section, we detail the data distribution of `PaperWritingBench`. As shown in Figure 5, Figure 6 and Table 10, ICLR 2025 papers exhibit higher visual and analytical density than CVPR 2025, averaging roughly twice as many figures (9.19 vs. 5.20) and tables (8.13 vs. 4.20).



(a) Citation Count Distribution (b) Material Length Distribution (c) Tables/Figures Distribution

Figure 6: ICLR 2025 Dataset Statistics

Metric	CVPR 2025	ICLR 2025
Visual Elements		
Number of Figures	5.20 ± 1.73	9.19 ± 5.39
Number of Tables	4.20 ± 1.65	8.13 ± 5.19
Citations		
Total Citations	58.52 ± 17.55	59.18 ± 20.01
P0 Citations (must-cite)	14.86 ± 6.72	13.65 ± 5.57
P1 Citations (good-to-cite)	43.66 ± 13.43	45.53 ± 17.36
Raw Material Word Counts		
Experimental Log ($\mathcal{E}$)	1529.91 ± 364.87	2386.71 ± 1484.56
Dense Idea Summary ($\mathcal{I}_{\text{Dense}}$)	1082.82 ± 263.45	1056.78 ± 165.62
Sparse Idea Summary ($\mathcal{I}_{\text{Sparse}}$)	591.42 ± 52.68	586.96 ± 53.81

Table 10: Detailed statistical breakdown of the `PaperWritingBench` dataset across CVPR 2025 and ICLR 2025. Values are reported as mean $\pm$ standard deviation.

Consequently, ICLR experimental logs are substantially longer (2,387 vs. 1,530 words). To ensure controlled evaluations, reverse-engineered input materials remain strictly consistent across both venues: Dense idea files average 1,083 (CVPR) and 1,057 (ICLR) words, while Sparse idea files are constrained to 591 (CVPR) and 587 (ICLR) words. Both datasets maintain comparable citation distributions, averaging $\sim$59 total references per manuscript.

D.2 Raw Material Extraction

The raw materials for `PaperWritingBench` are extracted using Gemini-3.1-Pro (prompts detailed in App. G.2). To systematically control the granularity of the provided concepts while ensuring the empirical data is extracted with high fidelity and minimal leakage, our pipeline employs two key mechanisms:

- **Controlled Idea Density (Dense vs. Sparse):** The *Dense* variant is prompted to explicitly preserve all mathematical formulations, loss functions, and LaTeX variables from the source text. Conversely, the *Sparse* variant abstracts away mathematical notation, describing formal architectures through high-level conceptual narratives.

- **Structured Context Injection:** Naive text-only extraction often corrupts tabular and mathematical data. To ensure the *Experimental Log* maintains strict fidelity to the original paper's content, we first parse the source PDF into markdown using MinerU (Wang et al., 2024). We independently extract visual assets and their original captions using PDFFigures 2.0 (Clark & Divvala, 2016). Rather than relying solely on unstructured markdown, we inject the extracted images of all tables and figures as multi-modal context blocks into Gemini-3.1-Pro. Sourced with this direct visual context during extraction, the model is then requested to flatten all explicit document references (e.g., stripping "as shown in Figure 2") and translate visual information into standalone factual observations, forcing downstream agents to autonomously reconstruct the empirical narrative.

D.3 Validity of Synthesized Input Materials

To confirm the reverse-engineered materials (*Idea Summaries* and *Experimental Logs*) represent realistic research artifacts without leaking ground-truth text, we conducted a dual evaluation. First, an ensemble of LLM judges (Gemini-3.1-Pro (The Gemini Team, 2026), GPT-5.4 (Singh et al., 2025), and Claude Sonnet 4.6 (Anthropic, 2026)) evaluated Technical Fidelity and Realism on a 1–5 scale. As Table 11 shows, the materials achieved high fidelity (∼4.74) and realism (∼4.51). Second, word-level lexical overlap (ROUGE-L (Lin, 2004)) against the ground-truth papers is exceptionally low (Table 12: 0.056 Sparse, 0.084 Dense), proving the benchmark supplies raw empirical facts rather than leaked narrative.

Model	Idea Type	Technical Fidelity (1–5)	Realism (1–5)
Gemini-3.1-Pro	Sparse / Dense	4.95 / 4.95	5.00 / 5.00
OpenAI GPT-5.4	Sparse / Dense	4.30 / 4.31	4.40 / 4.20
Claude Sonnet 4.6	Sparse / Dense	4.96 / 4.97	4.23 / 4.23

Table 11: Quality AutoEval Results on `PaperWritingBench` Raw Materials.

Idea Type	ROUGE-L Score
Sparse	0.0564
Dense	0.0840

Table 12: ROUGE-L Scores (`PaperWritingBench` Raw Materials vs. GT Paper).

D.4 Raw Material Example

In this section, we present a complete data sample comprising the sparse idea ($\mathcal{I}_{\text{sparse}}$), dense idea ($\mathcal{I}_{\text{dense}}$), and experimental log ($\mathcal{E}$) extracted from a randomly selected paper (Radman & Laaksonen, 2025) in the CVPR 2025 split.

Sparse Idea ($\mathcal{I}_{sparse}$)

Problem Statement
The Segment Anything Model (SAM) has established a new baseline for static image segmentation; however, it is structurally ill-equipped for Referring Audio-Visual Segmentation (Ref-AVS). Current foundation models like SAM suffer from two critical limitations in this context:

1. **Lack of Temporal Awareness:** SAM processes inputs as isolated static frames, failing to capture the temporal consistency and dynamic context necessary for video segmentation.
2. **Reliance on Explicit Interaction:** SAM depends on manual user prompts (points, boxes, or masks) to identify targets. It lacks the native ability to interpret implicit "multimodal prompts"—such as identifying an object described by a specific sound or textual description—without human intervention.

Core Hypothesis
We hypothesize that we can adapt the frozen, pre-trained knowledge of SAM for dynamic audio-visual scenes without heavy retraining by introducing two novel architectural components. First, a parallel **temporal modeling branch** can inject time-series context into the static encoder. Second, we can replace the manual "user-in-the-loop" prompting mechanism with **data-driven multimodal prompts**. By translating aligned audio and text features into the sparse and dense prompt formats that SAM natively understands, we believe the model can learn to "prompt itself" to segment temporally evolving objects described by complex semantic cues.

Proposed Methodology (High-Level Technical Approach)

We propose **TSAM**, a modified architecture that wraps the standard SAM backbone. Our approach focuses on minimal trainable additions while keeping the heavy image encoder frozen.

1. Temporal Modeling Branch (Context Injection)
Instead of retraining the heavyweight image encoder, we will introduce a lightweight auxiliary branch running in parallel.

- **Sequential Processing:** As the frozen encoder processes a frame, this parallel branch will accept features from current and previous steps.
- **Cached Memory & Adapters:** We will utilize a memory mechanism to store a running summary of previous frame representations. An adapter module will fuse these historical visual states with current audio-text embeddings. This allows the model to maintain object permanence and consistency across time.

2. Automated Multimodal Prompting
We aim to synthesize the "prompts" SAM expects (points and masks) using audio-visual-text correlations rather than manual clicks.

- **Sparse Prompting Module (Simulating Points):** We will design a query selection mechanism. By analyzing the correlation between the reference text and the audio stream, the system will select the most relevant "audio cues." These cues will be projected and treated as sparse queries (analogous to point prompts) to guide the mask decoder toward the sound source.
- **Dense Prompting Module (Simulating Masks):** We will implement a deep fusion module that uses cross-attention to mix visual features with aligned audio and text embeddings. This will generate a dense, spatial feature map that acts as a coarse "mask prompt," providing global context to the decoder.

3. Decoding
The final segmentation will be generated by the standard SAM mask decoder, which will be queried by our synthetic sparse and dense prompts, refined by the temporally aware features from our auxiliary branch.

Expected Contribution

- **Architectural Novelty:** A framework for extending static, unimodal foundation models (like SAM) into the temporal and multimodal domains without compromising their zero-shot capabilities.
- **Methodological Advancement:** A demonstration of "latent prompting," where semantic cues (audio/text) are mathematically translated into the geometric prompts (points/masks) required by segmentation models.
- **Theoretical Impact:** Establishing that lightweight temporal adapters and cross-modal attention mechanisms are sufficient to unlock video understanding in static image backbones.

Dense Idea ($\mathcal{I}_{dense}$)

Problem Statement
Referring Audio-Visual Segmentation (Ref-AVS) presents a complex challenge: segmenting objects in dynamic video scenes based on multimodal cues (audio, visual, and textual). Existing approaches, such as Video Object Segmentation (VOS) or Audio-Visual Segmentation (AVS), are limited by their reliance on single-modality cues or labor-intensive frame-level annotations. While Large Visual Foundation Models, specifically the Segment Anything Model (SAM), have revolutionized static image segmentation, they are fundamentally ill-suited for Ref-AVS in their current form. SAM lacks temporal awareness necessary for processing video frames and relies heavily on explicit user-interactive prompts (points, boxes) rather than interpreting the nuanced semantic

interactions between audio, text, and visual data. There is currently no architecture that effectively adapts SAM's zero-shot capabilities to the temporal and multimodal requirements of Ref-AVS without requiring explicit user guidance.

Core Hypothesis
We propose **TSAM (Temporal SAM)**, a novel architecture designed to extend SAM for dynamic audio-visual scenes. We hypothesize that by freezing the core SAM image encoder and augmenting it with a parallel **Temporal Modeling Branch**, we can inject spatio-temporal awareness into the model while preserving pre-trained knowledge. Furthermore, we propose replacing SAM's manual prompting mechanism with **data-driven multimodal prompts**. By synthesizing sparse queries (from audio-text alignment) and dense embeddings (from cross-modal attention), we can automate the prompting process, effectively guiding the mask decoder to segment targets defined by complex multimodal expressions.

Proposed Methodology (Detailed Technical Approach)

1. Input Formulation and Feature Extraction
We will process a sequence of inputs consisting of video frames, aligned audio, and reference text.

- **Visual Input:** n frames sampled at 1-second intervals, where each frame has a resolution of 1024×1024.
- **Audio Input:** Encoded offline using a pre-trained VGGish model to produce representations $\boldsymbol{F}_a \in \mathbb{R}^{n \times 128}$.
- **Text Input:** Extracted using a pre-trained RoBERTa model to generate representation $F_t \in \mathbb{R}^{l \times 768}$, where l is the sequence length.

To align these modalities with SAM, we will project both audio and text representations into the SAM visual embedding space with dimensionality d_{emb}. This yields audio cues $F_a \in \mathbb{R}^{n \times 1 \times d_{emb}}$ and text cues $F_t \in \mathbb{R}^{n \times l \times d_{emb}}$ (temporally expanded along n).

2. Temporal Modality Fusion and Cached Memory
We will implement a **Temporal Modality Fusion Layer (TMFL)** to synthesize expression-related multimodal cues. We define the fused cues as:

$$F'_a, F'_t = \mathrm{SA}(\mathrm{Concat}(F_a, F_t))$$

where $\mathrm{SA}(\cdot)$ applies temporal self-attention.
We will also utilize a **Cached Memory (CM)** mechanism. This will store $\hat{F}_a$, an accumulated summary of F'_a over time, capturing the mean modality cues and evolving audio context. Text cues are enriched via summation: $\hat{F}_t = F_t + F'_t$.

3. Temporal Modeling Branch
To address SAM's lack of temporal dynamics, we will introduce a trainable temporal modeling branch parallel to the final M blocks of the frozen SAM image encoder.

- **Structure:** The branch consists of blocks initialized with SAM's pre-trained weights.
- **Data Flow:** Each m-th temporal block ($m = 1, \ldots, M$) integrates the output of the previous temporal block y_{m-1}, the output of the corresponding frozen SAM block $y^{\mathrm{SAM}}_{(N-M)+(m-1)}$, and the enriched audio cues $\hat{F}_a$ via an Adapter Module (AM).
- **Formulation:**

$$x_m = y_{m-1} + \mathbf{CM}_m(y^{\mathrm{SAM}}_{(N-M)+(m-1)}) + \mathbf{AM}_m(\hat{F}_a)$$

  Here, $\mathbf{CM}_m$ is a cached memory applied to the SAM block output, and $\mathbf{AM}_m$ is a bottleneck adapter block facilitating deep interaction between audio, visual, and text modalities.

The final video cues F_v are derived by summing the output of the temporal branch (Y_M) and the final SAM encoder output (Y_N^{SAM}):

$$F_v = Y_M + Y_N^{SAM}$$

4. Multimodal Prompting Modules
We will replace SAM's manual prompts with two specific modules designed to query the mask decoder:

A. Sparse Prompting Module (SPM)
This module acts as a global context guide. We will employ a language-guided query selection mechanism to identify the k most relevant audio cues associated with the text expression. The indices of these cues, A_k, are selected via:

$$A_k = \text{Top}_k \left(\max \text{f}(\hat{F}_a \hat{F}_t^\top) \right)$$

where Top_k selects the top k indices and maxf computes the maximum along the cue dimension. These selected cues are fused with visual cues via cross-attention and mapped to sparse prompt embeddings.

B. Dense Prompting Module (DPM)
This module operates at a fine granularity. It generates spatially comprehensive embeddings through a sequence of cross-attention mechanisms:

1. **Audio Cross-Attention:** Aligns audio cues with visual cues.
2. **Text Cross-Attention:** Refines the result with text cues to focus on the specific target object.
3. **Refinement:** A feed-forward network processes the output to create dense embeddings for the SAM mask decoder.

5. Training Objective
The model will be trained end-to-end using a weighted sum of Binary Cross-Entropy ($\mathcal{L}_{\text{BCE}}$) and Intersection over Union ($\mathcal{L}_{\text{IoU}}$) losses to compare predicted masks against ground truth:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{IoU}}$$

We will set $\lambda = 1.0$ to balance the contributions of both loss components.

Expected Contribution

1. **Architecture:** A novel end-to-end framework (TSAM) that repurposes the Segment Anything Model for Referring Audio-Visual Segmentation without requiring extensive retraining of the image backbone.
2. **Temporal Adaptation:** The introduction of a lightweight Temporal Modeling Branch that enables SAM to capture intricate spatio-temporal interactions across video frames, overcoming its static-image limitation.
3. **Automated Multimodal Prompting:** A theoretical framework for converting audio-visual-text correlations into the sparse and dense prompts required by SAM, effectively replacing human interaction with data-driven multimodal guidance.

Experimental Log ($\mathcal{E}$)

1. Experimental Setup
We conducted a comprehensive evaluation of the proposed TSAM method for the Referring Audio-Visual Segmentation (Ref-AVS) task.

- **Datasets:**
  - **Ref-AVS Dataset:** We utilized the Ref-AVS dataset containing 20,000 text expressions and pixel-level annotations across 4,000 10-second videos.
  - **Object Categories:** The dataset included audible objects (20 musical instruments, 8 animals, 15 machines, 5 humans) and 3 categories of static, inaudible objects.
  - **Splits:**
    - *Training Set:* 2,908 videos.

    - *Validation Set:* 276 videos.
    - *Test Set:* 818 videos total, subdivided into:
        - *Seen:* 292 videos (categories present in training).
        - *Unseen:* 269 videos (categories not seen during training; tests generalization).
        - *Null:* 257 videos (text refers to non-existent/not visible objects; empty true masks).
- **Evaluation Metrics:**
  - **Standard Metrics:** We employed the Jaccard Index ($\mathcal{J}$) and F-score ($\mathcal{F}$) for the Seen and Unseen subsets.
  - **Null Subset Metric:** We employed the metric $\mathcal{S}$, which measures the ratio of the predicted mask area to the background area ($\mathcal{S} = \sqrt{\text{predicted mask area}/\text{background area}}$). A lower $\mathcal{S}$ value indicates better performance (less incorrect segmentation).
- **Baselines Compared:**
  - **AVS Methods (Augmented with Text):** AVSBench (PVT-v2 backbone), AVSegFormer (PVT-v2 backbone), GAVS (SAM backbone), SAMA (SAM backbone). *Note: We re-evaluated SAMA by integrating text with audio/visual inputs.*
  - **Ref-VOS Methods (Augmented with Audio):** ReferFormer (Video-Swin backbone), R2-VOS (Video-Swin backbone).
  - **Ref-AVS Method:** EEMC (Mask2Former backbone; previous state-of-the-art).
- **Implementation Details:**
  - **Model Initialization:** We initialized TSAM using the pre-trained ViT-B variant of SAM ($N = 12$, embedding dimension $d_{emb} = 256$).
  - **Architecture Settings:** The temporal branch consisted of $M = 4$ blocks. The number of selected audio queries was set to $k = 5$.
  - **Hardware:** Training was performed on eight AMD GPUs in a distributed setup.
  - **Optimizer:** AdamW optimizer was used.
  - **Hyperparameters:** Initial learning rate of $1 \cdot 10^{-4}$, batch size of 1.
  - **Training Duration:** The model was trained for 15 epochs with periodic evaluations.
  - **Model Selection:** We selected the best-performing model on the validation subset for testing.

2. Raw Numeric Data

Table 1: Performance comparison on the Ref-AVS dataset
*Note: Baselines marked with † had text integration added; baselines marked with ‡ had audio integration added; * marks our re-implementation of SAMA with text added.*

Method	Task	Visual Backbone	Seen J(%)	Seen F	Unseen J(%)	Unseen F	Null S(↓)
AVSBench	AVS†	PVT-v2	23.20	0.511	32.36	0.547	0.208
AVSegFormer	AVS†	PVT-v2	33.47	0.470	36.05	0.501	0.171
GAVS	AVS†	SAM	28.93	0.498	29.82	0.497	0.190
SAMA	AVS*	SAM	39.22	0.562	47.50	0.566	0.130
ReferFormer	Ref-VOS‡	V-Swin	31.31	0.501	30.40	0.488	0.176
R2-VOS	Ref-VOS‡	V-Swin	25.01	0.410	27.93	0.498	0.183
EEMC	Ref-AVS	M2F	34.20	0.513	49.54	0.648	**0.007**
TSAM (Ours)	Ref-AVS	SAM	**43.43**	**0.568**	**54.58**	**0.664**	0.017

Table 2: Ablation study on TSAM components and IoU loss
Legend: TB=Temporal Branch, TMFL=Temporal Modality Fusion Layer, DPM=Dense Prompting Module, SPM=Sparse Prompting Module, CM=Cached Memory, AM=Adapter Module. Mix(S+U) is the average of Seen and Unseen.

Setting	Seen J(%)	Seen F	Unseen J(%)	Unseen F	Mix(S+U) J(%)	Mix(S+U) F	Null S(↓)
(1) TSAM (Full)	**43.43**	0.568	**54.58**	**0.664**	**49.01**	**0.616**	0.017
(2) - TB	33.05	0.507	50.48	0.657	41.77	0.582	0.505
(3) - TMFL	40.35	0.579	45.54	0.627	42.95	0.603	0.018
(4) - DPM	42.72	0.580	49.10	0.647	45.91	0.614	0.018
(5) - SPM	43.04	0.580	49.75	0.652	46.40	**0.616**	0.018
(6) - SPM+DPM	42.60	0.602	40.58	0.604	41.59	0.603	0.018
(7) - CM(a+v)	42.07	0.544	49.11	0.659	45.59	0.602	0.018
(8) - CM(v)	42.75	0.549	51.18	0.660	46.97	0.605	0.018
(9) - AM	43.13	**0.600**	40.79	0.617	41.96	0.609	0.017
(10) - $\mathcal{L}_{\text{IoU}}$	38.29	0.564	42.15	0.631	40.22	0.598	**0.008**

Table 3: Effect of audio queries (k) and temporal branch depth (M)

Variation	Seen J(%)	Seen F	Unseen J(%)	Unseen F	Mix(S+U) J(%)	Mix(S+U) F	Null S(↓)
$k = 3$	43.58	**0.579**	50.43	0.655	47.01	**0.617**	0.018
$k = 5$	43.43	0.568	**54.58**	**0.664**	**49.01**	0.616	**0.017**
$k = 7$	43.24	0.573	46.95	0.630	45.10	0.601	0.018
$M = 2$	42.04	0.566	53.51	0.651	47.76	0.609	0.023
$M = 4$	**43.43**	0.568	**54.58**	**0.664**	**49.01**	**0.616**	**0.017**
$M = 6$	43.27	**0.575**	49.86	0.640	46.57	0.608	0.020

3. Qualitative Observations

Comparisons with State-of-the-Art:

- **Backbone Analysis:** We observed that methods utilizing prior segmentation visual backbones (SAM and Mask2Former) generally outperformed those based on PVT-v2 and V-Swin backbones.
- **SAM-Based Baseline Limitations:** Although GAVS and SAMA are SAM-based, they performed worse than EEMC. We noted that SAMA failed to leverage SAM's flexible, promptable nature, and GAVS lacked cohesive multimodal fusion.
- **TSAM Performance:** TSAM achieved the highest performance on both Seen and Unseen test sets. Specifically, TSAM improved over EEMC by 9.23% in Jaccard Index on the Seen set and 5.04% on the Unseen set.
- **Null Set Performance:** TSAM fell slightly behind EEMC on the Null test set (S value 0.017 vs 0.007). We attribute this to SAM's inherent limitation of always attempting to produce a segmentation mask even when no target object is present.

Ablation Observations:

- **Temporal Branch (TB):** Removing the temporal branch caused the most significant performance degradation (Seen Jaccard dropped from 43.43% to 33.05%). This confirmed the critical role of temporal modeling for generalizing across video frames.
- **Prompting Modules (SPM/DPM):** We found that the Sparse Prompting Module (SPM) and Dense Prompting Module (DPM) play complementary roles. Removing both simultaneously resulted in a notable decrease in segmentation performance.
- **Cached Memory (CM):** Omitting cached memory, especially the visual-only memory, degraded performance, highlighting the importance of shared memory for temporal alignment.
- **Adapter Module (AM):** The omission of the adapter module caused a significant performance drop, particularly on the Unseen test set, validating its role in facilitating deep multimodal interaction.
- **Loss Function:** Including the IoU loss ($\mathcal{L}_{\text{IoU}}$) improved performance on Seen and Unseen sets but slightly degraded the Null score.

Hyperparameter Sensitivity:

- **Audio Queries (k):** We found that $k = 5$ yielded optimal results. A lower value ($k = 3$) performed well on Seen data but worse on Unseen, while a higher value ($k = 7$) likely introduced irrelevant queries, overloading the decoder.

- **Temporal Depth (M):** $M = 4$ blocks provided the best balance. Shallower setups ($M = 2$) lacked sufficient temporal depth, while deeper setups ($M = 6$) appeared to introduce excessive complexity that impaired generalization.

Visual Qualitative Analysis:

- **Seen Test Set:** We observed that TSAM consistently produced high-quality masks for targeted objects. It successfully segmented inaudible objects when guided by textual cues (e.g., "The object behind the sounding women"), demonstrating effective processing of complex multimodal instructions.
- **Unseen Test Set:** TSAM demonstrated a remarkable capacity to generalize to novel objects. It effectively aligned audio-visual and textual inputs in novel scenes (e.g., segmenting a "truck moving" or "tuba being played" that were not in training categories).
- **Generalization:** The visual results underscored TSAM's ability to preserve SAM's pre-trained knowledge while using the temporal branch to understand dynamic scenes.

E Experiment Details

E.1 Models and Configuration

API Access We access all models from the Gemini-3 family (`gemini-3.1-pro-preview`, `gemini-3-flash-preview`, and `gemini-3-pro-image-preview`) via the Google Cloud Vertex AI platform. The GPT model (`gpt-5-2025-08-07`) is accessed via the official OpenAI API. Paper searches for verification and metadata fetching are conducted using the Semantic Scholar API (`https://api.semanticscholar.org/graph/v1/paper/search`).

Model Usage The primary manuscript writing backbone for PAPERORCHESTRA and all baselines is strictly fixed to Gemini-3.1-Pro. For literature discovery, PAPERORCHESTRA utilizes Gemini-3-Flash with Google Search grounding. To ensure a fair comparison, we also use Gemini-3-Flash for the citation gathering process in AI Scientist-v2.

Evaluation Settings For automated technical quality evaluations (AgentReview (Jin et al., 2024), AI Scientist-v2 Reviewer (Yamada et al., 2025), and ScholarPeer (Goyal et al., 2026)), we apply a universal temperature of 0.75 (Gemini-3.1-Pro) to balance instruction adherence with creative synthesis. For automated side-by-side evaluations, Gemini-3.1-Pro is set to a temperature of 0.0 to guarantee reproducibility. Since GPT-5 does not currently support temperature adjustment, we operate it at its fixed default temperature of 1.0 across all evaluations. During ScholarPeer (Goyal et al., 2026) evaluation, we disable the baseline scouting agent for all pipelines to isolate presentation quality from fixed experimental constraints.

Research Cutoff We align the research cutoff dates with the official submission deadlines of each venue: November 2024 for CVPR 2025 papers and October 2024 for ICLR 2025 papers.

E.2 Baseline Selection

We explain why several existing systems are not included as baselines:

- **OmniScientist** (Shao et al., 2025): Lacks a publicly available, reproducible codebase.
- **PaperRobot** (Wang et al., 2019) and **Data-to-Paper** (Technion-Kishony Lab, 2024): Do not support end-to-end generation of submission-ready LaTeX manuscripts.
- **AI-Researcher** (Tang et al., 2025a): Operates as a full research pipeline requiring highly structured intermediate states (e.g., pre-written sections, curated literature summaries)

Venue	Pipeline	Avg #Cites	Overall F1 (%)	P0 Recall (%)	P1 Recall (%)
CVPR 2025	AgentLaboratory	6.85	6.06	11.92	0.71
	GoogleAntigravity	10.14	19.44	40.49	2.38
	Ours (PAPERORCHESTRA)	**47.98**	**29.65**	**63.58**	**15.85**
ICLR 2025	AgentLaboratory	6.60	6.85	13.30	1.34
	GoogleAntigravity	8.60	16.73	34.19	2.27
	Ours (PAPERORCHESTRA)	**45.73**	**27.76**	**54.48**	**16.43**

Table 13: Citation F1 evaluation for additional baselines (evaluated via Gemini-3.1-Pro).

Comparison	Judge: Gemini-3.1-Pro			Judge: GPT-5		
	Win (%)	Tie (%)	Loss (%)	Win (%)	Tie (%)	Loss (%)
CVPR 2025						
Ours vs. AgentLaboratory						
Lit. Review Quality	100	0	0	96	4	0
Overall Paper Quality	78	22	0	56	33	11
Ours vs. GoogleAntigravity						
Lit. Review Quality	100	0	0	100	0	0
Overall Paper Quality	99	1	0	98	2	0
ICLR 2025						
Ours vs. AgentLaboratory						
Lit. Review Quality	99	1	0	100	0	0
Overall Paper Quality	62	35	3	58	34	8
Ours vs. GoogleAntigravity						
Lit. Review Quality	100	0	0	100	0	0
Overall Paper Quality	94	4	2	99	0	1

Table 14: Automated SxS evaluation results for PAPERORCHESTRA against additional baselines (AgentLaboratory, GoogleAntigravity) on CVPR 2025 and ICLR 2025.

rather than unconstrained raw materials. Adapting it to our setting would require reconstructing these abstractions, bypassing the core synthesis challenge our benchmark evaluates.

- **CycleResearcher** (Weng et al., 2025): Requires a structured BibTeX reference list as input—an artifact rarely available during early drafting. The model also fails on unstructured inputs outside its specific training format, preventing it from directly processing raw materials.
- **Generic RAG pipelines** ((Wu et al., 2025), (Go et al., 2026)): Designed primarily for literature surveys, these lack the specialized mechanisms required to autonomously draft, structure, and format full submission-ready research papers.

E.3 Additional Baselines

To confirm that PAPERORCHESTRA's gains stem from multi-agent orchestration rather than raw compute utilization or general tool access, we evaluated two additional baselines:

- **AgentLaboratory** (Schmidgall et al., 2025): A specialized paper solver scaled to 60–70 LLM calls to exactly match PAPERORCHESTRA's compute budget.
- **GoogleAntigravity** (Google, 2026): A strong general-purpose agentic system equipped with academic research skills (K-Dense-AI, 2026) and powered by the exact same Gemini-3.1-Pro backbone used in PAPERORCHESTRA.

As shown in Table 13, both baselines suffer from severe under-citation (averaging ∼7–10 references versus PAPERORCHESTRA's ∼46–48). In automated side-by-side (SxS) comparisons (Table 14), PAPERORCHESTRA dominates both baselines, confirming the necessity of specialized agents for deep citation coverage.

E.4 Citation Verification

To guarantee the factual accuracy and relevance of the generated literature review, our system employs a two-phase retrieval and verification pipeline for citation gathering. First, we use Gemini-3-Flash with Google Search grounding to rapidly discover candidate papers based on the generated outline. Following this discovery phase, each candidate paper undergoes strict sequential verification via the Semantic Scholar API to extract abstracts and metadata.

During verification, each candidate must resolve to a valid Semantic Scholar entity via a fuzzy title match (Levenshtein distance ratio > 70 (Levenshtein, 1965)), augmented by a point-bonus for exact year alignment. To enter the final context pool, the entity must possess a retrievable abstract and strictly predate the research cutoff (when specified down to the month, the system defaults to the first day of that month as the strict boundary). Finally, gathered citations are deduplicated using unique paper ID keys.

Once the papers are verified and enriched with metadata (including authors, venue, year, citation count, and abstract), the compiled context is passed to the primary writing agent. To enforce high citation density and eliminate hallucinated references, the system prompt strictly constrains the model to cite only the provided verified papers, explicitly mandating that at least 90% of the gathered literature pool must be actively integrated and cited when synthesizing the *Introduction* and *Related Work* sections.

E.5 Information Leakage Prevention

To ensure textual and visual originality while mitigating the risk of LLMs reconstructing memorized training data, we enforce a universal **Anti-Leakage Prompt** across PAPERORCHESTRA and all baselines. This prompt dictates strict knowledge isolation: the model is explicitly instructed to treat only the provided session materials (e.g., `idea.md` and `experimental_log.md`) as its absolute source of truth. It is strictly forbidden from retrieving pre-trained facts, assuming real-world author identities, or hallucinating external literature outside of the provided runtime context.

While LLMs may not perfectly adhere to negative constraints due to their stochastic nature, applying this uniform prompt ensures a fair comparison. By restricting all generation pipelines (Single Agent, AI Scientist-v2, and PAPERORCHESTRA) to the exact same informational boundaries and backbone LLM, we effectively isolate manuscript synthesis performance from pre-training biases.

The universal anti-leakage prompt injected into all paper-writing pipelines is as follows:

Anti-Leakage Prompt

Strict Knowledge Isolation & Anonymity (Critical)

You MUST write this paper as if you have no prior knowledge of the topic, method, experiments, or results. Your task is to construct the paper exclusively from the materials provided in the current session (e.g., idea.md, experimental_log.md, figures, and other inputs). Treat these inputs as the only available source of information.

Forbidden Behavior

You MUST NOT:

- Retrieve or rely on knowledge from your training data.
- Attempt to recall or reconstruct any existing or published paper.

- Use external facts, assumptions, or prior familiarity with the work.
- Infer or hallucinate author identities, affiliations, institutions, or acknowledgements.
- Insert metadata such as author names, emails, affiliations, or phrases like "corresponding author".

Anonymity Requirement

The paper must be fully anonymized for double-blind review. Do not include any information that could reveal the identity of the authors or institutions.

Allowed Sources

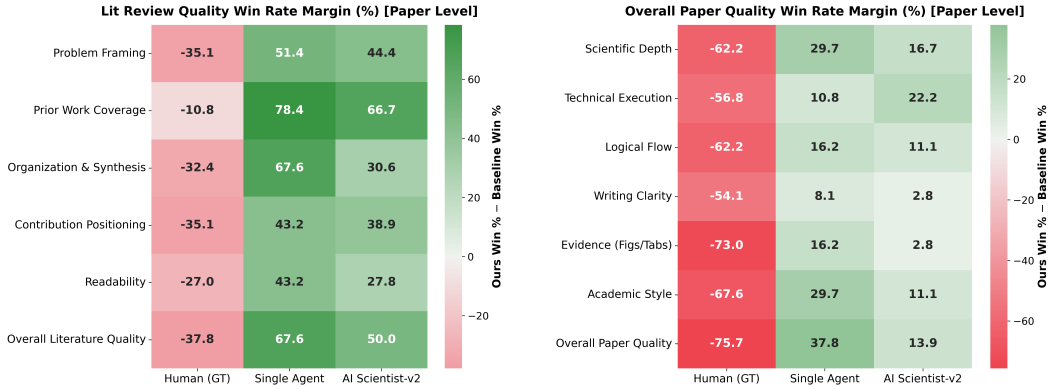
You may use only:

- The materials explicitly provided in this session.
- Logical reasoning derived from those materials.

Core Principle

The final paper must be an independent reconstruction derived solely from the provided inputs. This constraint is strict and overrides all other instructions.

E.6 Human Evaluation



(a) SxS Results: Literature Review Quality

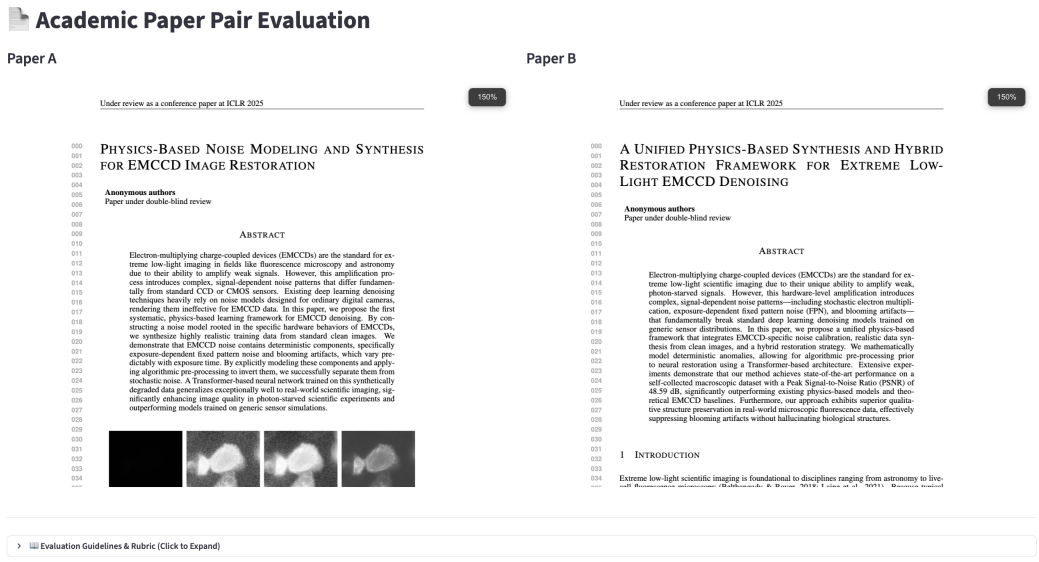
(b) SxS Results: Overall Paper Quality

Figure 7: **Human Side-by-Side (SxS) Evaluation Results.** Heatmaps display the win rate margin (%) of PAPERORCHESTRA against baselines across various evaluation dimensions. Positive values (green) favor our approach. PAPERORCHESTRA consistently outperforms both AI baselines (Single Agent and AI Scientist-v2), though a quality gap remains compared to the human-written ground truth (GT).

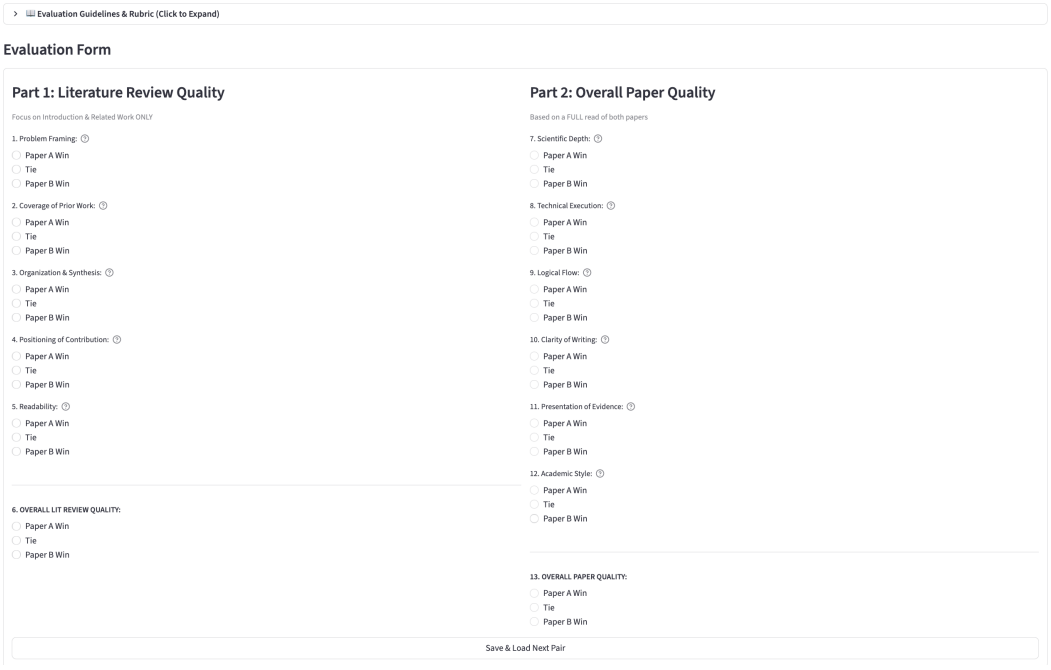
We developed a Streamlit interface (Figure 8) for blind SxS comparison of generated manuscripts. Annotators evaluated paper pairs on a 3-point scale (Win, Tie, Loss) across two aspects: (1) *Literature Review Quality* (evaluating the *Introduction* and *Related Work*) based on problem framing, prior work coverage, synthesis, positioning, and readability; and (2) *Overall Paper Quality* (evaluating the full manuscript) based on scientific depth, technical execution, logical flow, writing clarity, evidence presentation, and academic style.

Eleven AI researchers completed 180 randomized, paired evaluations across 40 sampled papers from `PaperWritingBench` (120 unique paper pairs). We observed Krippendorff's alpha scores of 0.229 for *Literature Review Quality* and 0.311 for *Overall Paper Quality*. These values align closely with the inherent subjectivity documented in expert peer-review experiments, such as the NeurIPS 2021 consistency study (Beygelzimer et al., 2021). To ensure evaluation rigor and prevent rater fatigue, individual workloads were strictly limited.

For each evaluation, annotators answered 12 fine-grained diagnostic questions before making a final holistic judgment. As shown in Figure 7, our approach, PAPERORCHESTRA, systematically outperforms all AI baselines across every evaluation metric. While a quality gap remains when compared to the human-written papers (GT), PAPERORCHESTRA achieves substantial win margins over all evaluated AI baselines in both literature review synthesis and overall paper quality.



(a) Side-by-Side Document View



(b) Evaluation Questionnaire

Figure 8: **Streamlit-based Human Evaluation Interface.** The custom UI displays two manuscripts side-by-side (a) and provides a fine-grained questionnaire alongside a final preference selection (b).

The detailed annotation guidelines and rubric provided to the human annotators are listed below:

Human Annotation Guidelines & Rubric

General Rules

- **Reading Scope:**
  - **Literature Review Quality** must be judged by reading the **Introduction and Related Work** sections *only*.
  - **Overall Paper Quality** must be judged based on a **full read** of both papers entirely.
- **Ignore Formatting and Metadata:** Some papers may include conference templates (e.g., CVF headers), watermarks, or text indicating prior publication. **Additionally, some papers may list author names while others are anonymized for review.** Please completely ignore these artifacts. Evaluations must be based purely on the intrinsic quality of the text and research, not the formatting, perceived prestige, or author identities.
- Evaluate each paper independently before comparing them.
- **Do not** base your decision solely on paper length or verbosity.

Literature Review Quality (Focus on Introduction & Related Work)

- **Motivation:** Does it clearly explain the problem, why it matters, and the gap in existing work?
- **Coverage:** Is the overview of prior research relevant and complete?
- **Synthesis:** Does it organize and group related work logically, rather than just blindly listing papers?
- **Positioning:** Does it clearly explain how *this* paper differs from existing methods?
- **Readability:** Is the text concise, clear, and easy to follow?

Overall Paper Quality (Holistic Review)

- **Scientific Depth:** Are the theoretical foundations, justifications, and experimental setups rigorous?
- **Execution:** Is the methodology implemented innovatively and effectively?
- **Logical Flow:** Does the paper transition smoothly from Abstract to Conclusion?
- **Clarity:** Is the writing precise and free of repetitive fluff or ambiguity?
- **Evidence:** Are figures, tables, and results cleanly integrated and referenced in the text?
- **Style:** Does it maintain a polished, consistent, professional academic tone?

F Paper Visualization

F.1 Side-by-Side (SxS) Paper Visualizations

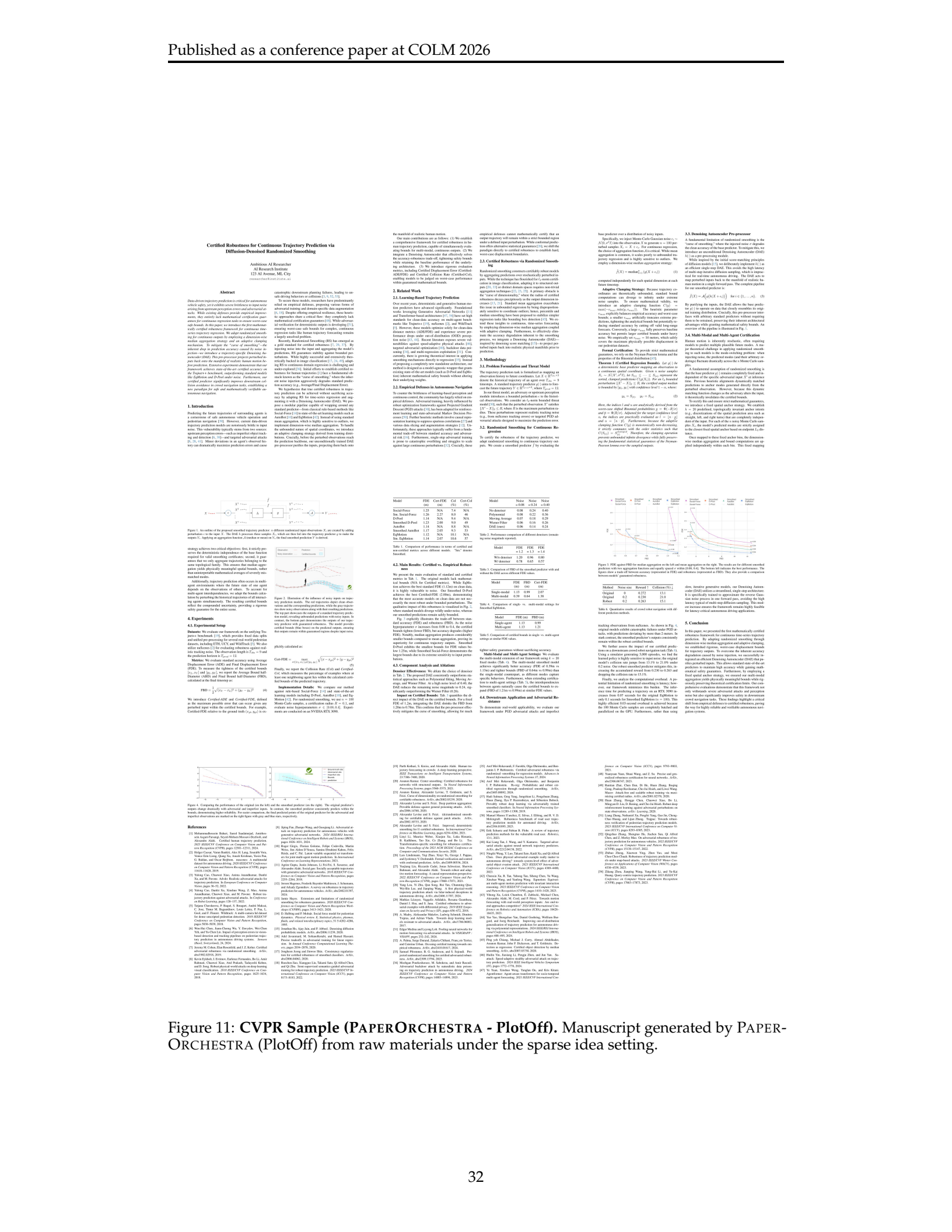
In this section, we present full manuscript generation results for two samples (one from CVPR (Bahari et al., 2025) and one from ICLR (Drakulić et al., 2025)) from `PaperWritingBench`. For each sample, we display the manuscripts generated from the original papers' raw materials under the sparse idea setting. We compare the final outputs produced by Single Agent, AI Scientist-v2, PAPERORCHESTRA (PlotOff), and PAPERORCHESTRA (PlotOn).



Figure 9: **CVPR Sample (Single Agent).** Manuscript generated by the Single Agent baseline from raw materials under the sparse idea setting.



Figure 10: **CVPR Sample (AI Scientist-v2).** Manuscript generated by the AI Scientist-v2 baseline from raw materials under the sparse idea setting.



Certified Robustness for Continuous Trajectory Prediction via Diffusion-Denoised Randomized Smoothing
Anthitious AI Researcher
AI Research Institute
123 AI Avenue, ML City
researcher@fakeinstitute.ai
Abstract
Data-driven trajectory prediction is critical for autonomous vehicle safety, yet it exhibits severe brittleness to input noise arising from sensor/gps-perception errors and adversarial attacks. While existing defenses provide empirical improvements, they entirely lack mathematical certification guarantees for continuous regression tasks to ensure worst-case safe bounds. In this paper, we introduce the first mathematically certified robustness framework for continuous time-series trajectory regression. We adapt randomized smoothing for continuous outputs by employing a dimension-wise median aggregation strategy and an adaptive clamping mechanism. To mitigate the "curse of smoothing"—the inherent drop in prediction accuracy caused by noise injection—we introduce a trajectory-specific Denoising Autoencoder (DAE). This pre-processor projects perturbed inputs back onto the manifold of realistic human motion before prediction. Extensive experiments demonstrate that our framework achieves state-of-the-art certified accuracy on the Trajnet++ benchmark, outperforming standard models like TgNMotion and D-Pool under noise. Furthermore, our certified predictor significantly improves downstream collision avoidance in crowd navigation tasks, establishing a new paradigm for safe and mathematically verifiable autonomous navigation.
1. Introduction
Predicting the future trajectories of surrounding agents is a cornerstone of safe autonomous vehicle operation and pedestrian navigation [16]. However, current data-driven trajectory prediction models are notoriously brittle to input noise. This vulnerability typically stems from two sources: upstream perception errors—such as imperfect object tracking and detection [6, 16]—and targeted adversarial attacks [8, 28, 41]. Minor deviations in an agent's observed history can dramatically maximize prediction errors and cause catastrophic downstream planning failures, leading to unsafe driving behaviors or collisions [5, 9, 52, 53].
To secure these models, researchers have predominantly relied on empirical defenses, proposing various forms of adversarial training and domain-specific data augmentation [6, 10]. Despite offering empirical resilience, these heuristic approaches share a critical flaw: they completely lack mathematical certification guarantees [48]. While adversarial verification for deterministic outputs is developing [13], ensuring worst-case safe bounds for complex, continuous regression tasks like human trajectory forecasting remains a largely unsolved problem.
Recently, Randomized Smoothing (RS) has emerged as a gold standard for certified robustness [7, 29, 37]. By injecting noise into the input and aggregating the model's predictions, RS guarantees stability against bounded perturbations. While highly successful and computationally theoretically backed in image classification [17, 24, 49], adapting RS to continuous domain regression is challenging and under-explored [34]. Initial efforts to establish certified robustness for human trajectories [1] face a fundamental obstacle known as the "curse of smoothing," where the inherent noise injection aggressively degrades standard prediction accuracy (e.g., Average/Final Displacement Error).
We hypothesize that true certified robustness in trajectory prediction can be achieved without sacrificing accuracy by adapting RS for time-series regression and augmenting it with a Denoising Autoencoder (DAE). We propose a modular pipeline capable of wrapping around any standard predictor—from classical rule-based methods like Social-Force [14] to state-of-the-art learning models such as AutoBot [10] and EgoMotion [42]. Instead of using standard mean aggregation, which is highly sensitive to outliers, we implement dimension-wise median aggregation. To handle the unbounded nature of spatial coordinates, we introduce an adaptive clamping strategy derived from training distributions. Crucially, before the perturbed observations reach the prediction backbone, our unconditionally trained DAE pre-processor purifies the inputs, projecting them back onto the manifold of realistic human motion.
Our main contributions are as follows: (1) We establish a comprehensive framework for certified robustness in human trajectory prediction, capable of simultaneously evaluating bounds for multi-modal, continuous outputs. (2) We integrate a Denoising Autoencoder that effectively solves the accuracy-robustness trade-off, tightening safety bounds while retaining the baseline performance of the underlying architecture. (3) We introduce rigorous evaluation metrics, including Certified Displacement Error (Certified-ADE/FDE) and Certified Collision Rate (Cert-CoI), enabling models to be judged on worst-case performance within guaranteed mathematical bounds.
2. Related Work
2.1. Learning-Based Trajectory Prediction
Over recent years, deterministic and generative human motion predictors have advanced significantly. Foundational works leveraging Generative Adversarial Networks [11] and Transformer-based architectures [47, 54] have set high standards for clean-data accuracy on multi-agent benchmarks like Trajnet++ [18], nuScenes [2], and WildTrack [1]. However, these models optimize solely for clean-data distance metrics (ADE/FDE) and experience severe performance drops under out-of-distribution (OOD) perception noise [43, 44]. Recent literature exposes severe vulnerabilities against speed-adaptive physical attacks [46], targeted adversarial optimization [45], backdoor data poisoning [38], and multi-regression exploitation [31]. Concurrently, there is growing theoretical interest in applying smoothing mechanisms directly to regression [35]. Instead of proposing a complex new standalone architecture, our method is designed as a model-agnostic wrapper that grants existing state-of-the-art models (such as D-Pool and EgoMotion) inherent mathematical safety bounds without altering their underlying weights.
2.2. Empirical Defenses in Autonomous Navigation
To counter the brittleness of learning-based perception and continuous control, the community has largely relied on empirical defenses. Adversarial training, heavily influenced by robust optimization frameworks against Projected Gradient Descent (PGD) attacks [30], has been adapted for reinforcement learning and state-adversarial Markov Decision Processes [50]. Further heuristic methods involve case-based regression learning to suppress spurious correlations [27] and various data slicing and augmentation strategies [12]. Unfortunately, these approaches typically suffer from a fundamental trade-off between standard accuracy and adversarial risk [51]. Furthermore, single-step adversarial training is prone to catastrophic overfitting and struggles to scale against large continuous perturbations [32]. Crucially, these empirical defenses cannot mathematically certify that an output trajectory will remain within a strict bounded region under a defined input perturbation. While conformal prediction offers alternative statistical guarantees [26], we shift the paradigm directly to certified robustness to establish hard, worst-case displacement boundaries.
2.3. Certified Robustness via Randomized Smoothing
Randomized smoothing constructs certifiably robust models by aggregating predictions over stochastically perturbed inputs. While the technique has flourished for ℓ_2-norm certification in image classification, adapting it to structured outputs [20, 33] or distinct domain spaces requires non-trivial aggregation techniques [22, 23, 25]. A primary obstacle is the "curse of dimensionality," where the radius of certified robustness decays precipitously as the output dimension increases [15, 21]. Standard mean aggregation exacerbates this issue in unbounded regression by being disproportionately sensitive to coordinate outliers; hence, percentile and median smoothing have been proposed to stabilize simpler regression tasks like bounding box detection [45]. We extend these insights to continuous, time-series forecasting by employing dimension-wise median aggregation coupled with adaptive clamping. Furthermore, to effectively eliminate the accuracy degradation inherent to the smoothing process, we integrate a Denoising Autoencoder (DAE)—inspired by denoising score matching [13]—to project perturbed inputs back into realistic physical manifolds prior to prediction.
3. Methodology
3.1. Problem Formulation and Threat Model
The trajectory prediction task is formalized as mapping an observation history to future coordinates. Let $X \in \mathbb{R}^{T_{obs} \times 2}$ denote the historical trajectory of an agent over $T_{obs} = 9$ timesteps. A standard trajectory predictor $g(\cdot)$ aims to forecast the future trajectory $Y \in \mathbb{R}^{T_{pred} \times 2}$, where $T_{pred} = 12$.
In our threat model, an adversary or upstream perception module introduces a bounded perturbation ϵ to the historical observations. We consider an ℓ_2-norm bounded threat model [30], such that the perturbed observation X' satisfies $\|X' - X\|_2 \leq R$, where R is the maximum perturbation radius. These perturbations represent realistic tracking noise (e.g., from occlusions tracking errors) or targeted PGD adversarial attacks designed to maximize the prediction error.
3.2. Randomized Smoothing for Continuous Regression
To certify the robustness of the trajectory predictor, we adapt randomized smoothing to continuous trajectory outputs. We create a smoothed predictor f by evaluating the base predictor over a distribution of noisy inputs.
Specifically, we inject Monte-Carlo Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$ into the observation X to generate $n = 100$ perturbed samples $X_i = X + \epsilon_i$. For continuous regression, the choice of aggregation function A is critical. While mean aggregation is common, it scales poorly to unbounded trajectory regression and is highly sensitive to outliers. We employ a dimension-wise *median aggregation strategy*:
$$\bar{f}(X) = \text{median}_{i=1}^{n}(g(X + \epsilon_i)) \tag{1}$$
computed independently for each spatial dimension at each future timestep.
Adaptive Clamping Strategy: Because trajectory coordinates are theoretically unbounded, standard bound computations can diverge to infinity under extreme noise samples. To ensure mathematical validity, we introduce an adaptive clamping function $C(y) = \max(-v_{max}, \min(y, v_{max}))$. The heuristic parameter v_{max} explicitly balances empirical accuracy and worst-case bounds: a smaller v_{max} artificially truncates extreme predictions, tightening the analytical bounds but potentially reducing standard accuracy by cutting off valid long-range forecasts. Conversely, a large v_{max} fully preserves baseline accuracy but permits larger certified bounds under heavy noise. We empirically set $v_{max} = 10$ meters, which safely covers the maximum physically possible displacement in our pedestrian datasets.
Formal Certification: To provide strict mathematical guarantees, we rely on the Neyman-Pearson lemma and the properties of the Binomial distribution [40].
Theorem 1 (Certified Regression Bounds). *Let $g(\cdot)$ be a deterministic base predictor mapping an observation to a continuous spatial coordinate. Given n noise samples $X_i \sim \mathcal{N}(X', \sigma^2 I)$, let $S_{(1)} \leq \ldots \leq S_{(n)}$ represent the sorted, clamped predictions $C(g(X_i))$. For an ℓ_2-bounded perturbation $\|X' - X\|_2 \leq R$, the median output prediction is bounded by $[y_{\downarrow}, y_{\uparrow}]$ with confidence level $1 - \alpha$, where:*
$$y_{\downarrow} = S_{(l)}, \quad y_{\uparrow} = S_{(u)} \tag{2}$$
Here, the indices l and u are analytically derived from the worst-case shifted Binomial probabilities $p = \Phi(-R/\sigma)$ and $\bar{p} = \Phi(R/\sigma)$. Adjusted for the target confidence level α, the indices are practically evaluated as $l = \lfloor n \cdot \underline{p} \rfloor$ and $u = \lceil n \cdot \bar{p} \rceil$. Furthermore, because the adaptive clamping function $C(y)$ is monotonically non-decreasing, it strictly commutes with the order statistics such that $C(S_{(i)}) = S^{(C)}_{(i)}$. Therefore, the clamping operation prevents unbounded infinite divergence while fully preserving the fundamental statistical guarantees of the Neyman-Pearson lemma over the sampled outputs.
3.3. Denoising Autoencoder Pre-processor
A fundamental limitation of randomized smoothing is the "curse of smoothing," where the injected noise σ degrades the clean accuracy of the base predictor. To mitigate this, we introduce an unconditional Denoising Autoencoder (DAE) $h(\cdot)$ as a pre-processing module.
While inspired by the initial score-matching principles of diffusion models [13], we deliberately implement $h(\cdot)$ as an efficient single-step DAE. This avoids the high latency of multi-step iterative diffusion sampling, which is impractical for real-time autonomous driving. The DAE acts to map perturbed inputs back to the manifold of realistic human motion in a single forward pass. The complete pipeline for our smoothed predictor is:
$$\bar{f}(X) = \mathcal{A}\left(g(h(X + \epsilon_i))\right) \quad \text{for } i \in \{1, \ldots, n\}. \tag{3}$$
By purifying the inputs, the DAE allows the base predictor $g(\cdot)$ to operate on data that *closely* resembles its original training distribution. Crucially, this pre-processor interfaces with arbitrary standard predictors without requiring them to be retrained, preserving their inherent architectural advantages while granting mathematical safety bounds. An overview of the pipeline is illustrated in Fig. 1.
3.4. Multi-Modal and Multi-Agent Certification
Human motion is inherently stochastic, often requiring models to predict multiple plausible future modes. A natural theoretical challenge in applying randomized smoothing to such models is the mode-switching problem: when injecting noise, the predicted modes (and their arbitrary ordering) fluctuate drastically across the n Monte Carlo samples.
A fundamental assumption of randomized smoothing is that the base predictor $g(\cdot)$ remains completely fixed and independent of the specific adversarial input. N at inference time. Previous heuristic alignments dynamically matched predictions to anchor modes generated directly from the perturbed observation. However, because this dynamic matching functions changes as the adversary alters the input, it theoretically invalidates the certified bounds.
To rectify this and ensure strict mathematical guarantees, we introduce a fixed spatial anchor strategy. We establish $k = 20$ predefined, topologically invariant anchor intents (e.g., discretizations of the spatial prediction area such as straight, left, and right turns) that are completely independent of the input. For each of the n noisy Monte Carlo samples X_i, the model's predicted modes are strictly assigned to the closest fixed spatial anchor based on endpoint L_2 distance.
Once mapped to these fixed anchor bins, the dimension-wise median aggregation and bound computations are applied independently within each bin. This fixed mapping

Figure 1. An outline of the proposed smoothed trajectory predictor: n different randomized input observations X_i are created by adding perturbations ϵ to the input X. The DAE h processes these samples X_i, which are then fed into the trajectory predictor g to make the outputs Y_i. Applying an aggregation function A (median or mean) on Y_i, the final smoothed prediction Y is derived.

strategy achieves two critical objectives: first, it strictly preserves the deterministic independence of the base function required for valid smoothing certificates; second, it guarantees that we only aggregate trajectories belonging to the same topological family. This ensures that median aggregation yields physically meaningful spatial bounds, rather than uninterpretable mathematical averages of severely mismatched modes.
Additionally, trajectory prediction often occurs in multi-agent environments where the future state of one agent depends on the observations of others. To account for multi-agent interdependencies, we adapt the bounds calculation by partitioning the historical trajectories of all interacting agents simultaneously. The resulting certified bounds reflect the compounded uncertainty, providing a rigorous safety guarantee for the entire scene.
4. Experiments
4.1. Experimental Setup
Datasets: We evaluate our framework on the unifying Trajectron++ benchmark [18], which provides fixed data splits and unified pre-processing for several real-world pedestrian datasets, including ETH, UCY, and WildTrack [1]. We also utilize nuScenes [2] for evaluating robustness against real-world tracking noise. The observation length is $T_{obs} = 9$ and the prediction horizon is $T_{pred} = 12$.
Metrics: We evaluate standard accuracy using Average Displacement Error (ADE) and Final Displacement Error (FDE). To measure the tightness of the certified bounds $[x_{\downarrow}, x_{\uparrow}]$ and $[y_{\downarrow}, y_{\uparrow}]$, we report the Average Bound half-Diameter (ABD) and Final Bound half-Diameter (FBD), calculated at the final timestep as:
$$\text{FBD} = \frac{1}{2}\sqrt{(x_{\uparrow} - x_{\downarrow})^2 + (y_{\uparrow} - y_{\downarrow})^2} \tag{4}$$
We introduce *Certified-ADE* and *Certified-FDE*, defined as the maximum possible error that can occur given any perturbed input within the certified bounds. For example, Certified-FDE relative to the ground truth (x_{gt}, y_{gt}) is explicitly calculated as:
$$\text{Cert-FDE} = \max_{x \in [x_{\downarrow}, x_{\uparrow}], y \in [y_{\downarrow}, y_{\uparrow}]} \sqrt{(x - x_{gt})^2 + (y - y_{gt})^2} \tag{5}$$
Finally, we report the Collision Rate (CoI) and *Certified CoI*, which measures the percentage of examples where at least one neighboring agent lies within the calculated certified bounds of the predicted trajectory.
Implementation Details: We compare our method against rule-based Social-Force [14] and state-of-the-art learning models including D-Pool, AutoBot [10], and EgoMotion [42]. For randomized smoothing, we use $n = 100$ Monte-Carlo samples, a certification radius $R = 0.1$, and evaluate noise hyperparameters $\sigma \in \{0.05, 0.4\}$. Experiments are conducted on an NVIDIA RTX 3090.

Figure 2. Illustration of the influence of noisy inputs on trajectory prediction models. The red trajectories display clean observations and the corresponding predictions, while the gray trajectories show noisy observations along with their resulting predictions. The top part showcases the outputs of a standard trajectory predictor, highlighting its susceptibility to input perturbations. In contrast, the bottom part demonstrates the outputs of our trajectory predictor with guaranteed robustness. The model provides certified bounds (blue boxes) on the predicted outputs, ensuring that outputs remain within guaranteed regions despite input noise.

Model	FDE (m)	Cert-FDE (m)	CoI (%)	Cert-CoI (%)
Social-Force	1.25	N/A	7.4	N/A
Soc. Social-Force	1.26	2.27	8.0	46
D-Pool	1.24	N/A	9.4	N/A
Smoothed D-Pool	1.23	2.00	9.0	49
AutoBot	1.17	2.05	8.3	N/A
Smoothed AutoBot	1.17	2.05	8.3	53
EgoMotion	1.12	N/A	10.1	N/A
Soc. EgoMotion	1.14	2.07	10.6	57

Table 1. Comparison of performance in terms of certified and non-certified metrics across different models. "Soc." denotes Smoothed.

Model	Noise σ = 0.08	Noise σ = 0.24	Noise σ = 0.40
No-denoiser	0.08	0.24	0.40
Polynomial	0.08	0.22	0.36
Moving Average	0.07	0.18	0.28
Wiener Filter	0.06	0.16	0.28
DAE (ours)	0.06	0.14	0.24

Table 2. Performance comparison of different denoisers (remaining noise magnitude reported).

Model	FDE $\sigma = 1.2$	FDE $\sigma = 1.3$	FDE $\sigma = 1.4$
W/o denoiser	1.20	0.96	0.80
W/ denoiser	0.78	0.65	0.57

Table 3. Comparison of FBD of the smoothed predictor with and without the DAE across different FDE values.

Model	FDE (m)	FBD (m)	Cert-FDE (m)
Single-modal	1.13	0.99	2.07
Multi-modal	0.39	0.64	1.38

Table 4. Comparison of single- vs. multi-modal settings for Smoothed EgoMotion.

Model	FDE (m)	FBD (m)
Single-agent	1.13	0.99
Multi-agent	1.13	1.21

Table 5. Comparison of certified bounds in single- vs. multi-agent settings at similar FDE values.

Figure 3. FDE against FBD for median aggregation on the left and mean aggregation on the right. The results are for different smoothed predictors with two aggregation functions and equally spaced σ within [0.00, 0.4]. The bottom left indicates the best performance. The figures show a trade-off between accuracy (represented in FDE) and robustness (represented as FBD). They also provide a comparison between models' guaranteed robustness.

Method	Noise size	Reward ↑	Collision (%) ↓
Original	0	0.232	13.1
Original	0.2	0.230	21.0
Robust	0.2	0.283	15.1

Table 6. Quantitative results of crowd robot navigation with different prediction methods.

tracking observations from nuScenes. As shown in Fig. 5, original models exhibit catastrophic failures under PGD attacks, with predictions deviating by more than 2 meters. In stark contrast, the smoothed predictor's outputs consistently remain within the certified bounds.
We further assess the impact of our certified predictors on a downstream crowd-robot navigation task (Tab. 6). Using a simulator generating 5,000 episodes, we find the learned policy is highly sensitive to input noise: the original model's collision rate jumps from 13.1% to 21.0% under 0.2 noise. Our robust smoothed predictor mitigates this, increasing the accumulated reward from 0.230 to 0.283 and dropping the collision rate to 15.1%.
Finally, we analyze the computational overhead. A potential limitation of randomized smoothing is latency; however, our framework minimizes this burden. The inference time for predicting a trajectory on an RTX 3090 increases from 0.07 seconds for the original EgoMotion to only 0.1 seconds for Smoothed EgoMotion ($n = 100$). This highly efficient 0.03-second overhead is achieved because the 100 Monte Carlo samples are completely batched and parallelized on the GPU. Furthermore, rather than using slow, iterative generative models, our Denoising Autoencoder (DAE) utilizes a streamlined, single-step architecture. It is specifically trained to approximate the inverse Gaussian noise process in one forward pass, avoiding the high latency typical of multi-step diffusion sampling. The modest increase ensures the framework remains highly feasible for latency-critical autonomous driving applications.
5. Conclusion
In this paper, we presented the first mathematically certified robustness framework for continuous time-series trajectory prediction. By adapting randomized smoothing through dimension-wise median aggregation and adaptive clamping, we established rigorous, worst-case displacement bounds for trajectory outputs. To overcome the inherent accuracy degradation caused by noise injection, we successfully integrated an efficient Denoising Autoencoder (DAE) that purifies perturbed inputs. This allows standard state-of-the-art predictors to maintain high accuracy while gaining mathematical safety guarantees. Furthermore, by employing a fixed spatial anchor strategy, we ensured our multi-modal aggregation yields physically meaningful bounds while rigorously preserving theoretical certification limits. Our comprehensive evaluations demonstrate that this framework not only withstands severe adversarial attacks and perception noise but also significantly improves safety in downstream crowd navigation tasks. These findings highlight a critical shift from empirical defenses to certified robustness, paving the way for highly reliable and verifiable autonomous navigation systems.

Figure 4. Comparing the performance of the original (on the left) and the smoothed predictor (on the right). The original predictor's outputs change drastically with adversarial and imperfect inputs. In contrast, the smoothed predictor consistently predicts within the bounds, demonstrating higher reliability. For easier comparison, the final predicted positions of the original predictor for the adversarial and imperfect observations are marked on the right figure with gray and blue stars, respectively.

References

[1] Mohammadreza Baharani, Sumit Sudarsanjal, Anirbesh Chattopadhyay, Ramtin Tabatabaei, Steven Tuerk, and Alexander Aiken. Certified human trajectory prediction. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12301–12311, 2024.
[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, G. Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2019.
[3] Yuhong Cao, Chaowei Xiao, Anima Anandkumar, Danfei Xu, and M. Pavone. Advdo: Realistic adversarial attacks for trajectory prediction. *In European Conference on Computer Vision*, pages 36–52, 2022.
[4] Yuhong Cao, Danfei Xu, Xinshuo Wang, Z. Mao, Anima Anandkumar, Xinshuo Wang, and M. Pavone. Robust trajectory prediction against adversarial attacks. *In Conference on Robot Learning*, pages 128–137, 2022.
[5] Tatjana Chavdarova, P. Baqué, S. Bouquet, Andrii Maksai, C. Jose, Timur M. Bagautdinov, Louis Lettry, P. Fua, L. Gool, and F. Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018.
[6] Wen-Hsi Chen, Jiann-Cherng Wu, Y. Danyelee, Wei-Chen Yeh, and Yu-Chen Lin. Impact of perception errors in vision-based detection and tracking pipelines on pedestrian trajectory prediction in autonomous driving systems. *Sensors (Basel, Switzerland)*, 24, 2024.
[7] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. *ArXiv*, abs/1902.02918, 2019.
[8] Kevin Eykholt, I. Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.
[9] Jiping Fan, Zhenpo Wang, and Guoqing Li. Adversarial attack on trajectory prediction for autonomous vehicles with generative adversarial networks. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1026–1031, 2024.
[10] Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and C. Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. *International Conference on Learning Representations*, 2021.
[11] Agrim Gupta, Justin Johnson, Li Fei-Fei, S. Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.
[12] Jayson Hagness, Frederik Bayrakar Mathiesen, J. Scheiman, and Arkady Zgonnikov. A survey on robustness in trajectory prediction for autonomous vehicles. *ArXiv*, abs/2402.01507, 2024.
[13] Jamie Heyes. Extensions and limitations of randomized smoothing for robustness guarantees. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 3413–3421, 2020.
[14] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, 51 5:4282–4286, 1995.
[15] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.
[16] Adel Javanmard, M. Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. *In Annual Conference Computational Learning Theory*, pages 2034–2078, 2020.
[17] Aounon Kumar and T. Goldstein. Center smoothing: Certified robustness for networks with structured outputs. *In Neural Information Processing Systems*, 2021.
[18] Agrim Gupta, Justin Johnson, Li Fei-Fei, S. Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.
[19] Parth Kothari, S. Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 23:7386–7400, 2020.
[20] Aounon Kumar. Center smoothing: Certified robustness for networks with structured outputs. *In Neural Information Processing Systems*, pages 5560–5575, 2021.
[21] Aounon Kumar, Alexander Levine, T. Goldstein, and S. Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. *ArXiv*, abs/2002.03239, 2020.
[22] Alexander Levine and S. Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. *ArXiv*, abs/2006.14768, 2020.
[23] Alexander Levine and S. Feizi. (de)randomized smoothing for certifiable defense against patch attacks. *ArXiv*, abs/2002.10733, 2020.
[24] Alexander Levine and S. Feizi. Improved, deterministic smoothing for l1 certified robustness. *In International Conference on Machine Learning*, pages 6254–6264, 2021.
[25] Linyi Li, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1289–1310, 2020.
[26] Lars Lindemann, Yiqi Zhao, Xinyi Yu, George J. Pappas, and Jyotirmoy V. Deshmukh. Formal verification and control with conformal prediction. *ArXiv*, abs/2409.00536, 2024.
[27] Yaojiang Liu, Riccardo Cadei, Jonas Schweisthal, Sherwin Bahmani, and Alexander Alahi. Towards robust and adversarial motion forecasting: A causal representation perspective. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17067–17071, 2021.
[28] Yang Liu, Yi Zhu, Qun Song, Rui Tan, Chunming Qiao, Wei-Bin Lee, and Jianping Wang. A fine-grained physical trajectory prediction attack via false-induced deceptions in autonomous driving. *ArXiv*, abs/2406.11787, 2024.
[29] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel J. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672, 2018.
[30] A. Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2017.
[31] Edgar Medina and Leyong Loh. Profiling neural networks for motion forecasting via adversarial attacks. *In VISIGRAPP , VISAPP*, pages 232–242, 2024.
[32] A. Palma, Serge Durand, Zakaria Chihani, Franç ois Terrier, and Caterina Urban. On using certified training towards empirical robustness. *ArXiv*, abs/2410.01617, 2024.
[33] Ji Pei, Wenlong Deng, Junyang Wang, and S. Kaushik. Provably robust conformal prediction with improved efficiency. *ArXiv*, abs/2306.03580, 2024.
[34] Huaxiu Yao, Xiaoqing Lin, Tatsuo Sato, Qi-Alfred Chen, and Qi Zhu. Semi-supervised semantics-guided adversarial training for robust trajectory prediction. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8173–8182, 2023.
[35] Aref Miri Rekavandi, F. Farokhi, Olga Ohrimenko, and Benjamin I. P. Rubinstein. Rs-reg: Probabilistic and robust certified regression through randomized smoothing. *ArXiv*, abs/2405.04602, 2024.
[36] Haonan Sheng, Dong Tian, Jiangtian Li, Pengzhuo Zhang, Hanbo Zhang, Dan Pi Ruzenkova, and Sébastien Rubeck. Provably robust deep learning via adversarially trained smoothed classifiers. *In Neural Information Processing Systems*, pages 11289–11300, 2019.
[37] Manuel Munoz Sanchez, E. Silvas, J. Ebbing, and R. V. D. Molengraft. Robustness benchmark of road user trajectory prediction models for automated driving. *ArXiv*, abs/2304.01895, 2023.
[38] Erik Schaffernicht and Fabian B. Flohr. A review of trajectory prediction methods for the vulnerable road user. *Robotics*, 13:1, 2023.
[39] Kai Liang Tan, J. Wang, and Y. Kantaros. Targeted adversarial attacks against neural network trajectory predictors. *ArXiv*, abs/2212.04138, 2022.
[40] Yuxiao Wang, Yu Yan, Sirui Cai, Xiaolei Xun, Kaidi Xu, and Qi-Alfred Chen. Does physical adversarial example really matter to autonomous driving? towards system-level effect of adversarial object evasion attack. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4389–4400, 2023.
[41] Chaowei Xiao, R. Tan, Yuhong Tan, Sibeng Chen, Ya Wang, Xinchao Wang, and Yunfeng Wang. Egomotion: Egocentric multi-agent motion prediction with invariant interaction encoding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1410–1420, 2021.
[42] Yuhong Na, Louis Chambon, E. Zablocki, Mickael Chen, Alexandre Alahi, M. Cord, and P. Pérez. Towards motion forecasting with real-world perception inputs: Are end-to-end approaches competitive? *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18426–18433, 2023.
[43] Yue Yao, Shengchao Yan, Daniel Goehring, Wolfram Burgard, and Joerg Reichardt. Improving out-of-distribution generalization of trajectory prediction for autonomous driving via polynomial representations. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 488–495, 2024.
[44] Ping yeh Chiang, Michael J. Curry, Ahmed Abdelkader, Aounon Kumar, John P. Dickerson, and T. Goldstein. Detection as regression: Certified object detection by median smoothing. *ArXiv*, abs/2007.13761, 2020.
[45] Haibo Yin, Jiarong Li, Penggo Zhou, and Jun Yan. Sp-attack: Speed-adaptive stealthy adversarial attacks on trajectory prediction. *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1772–1778, 2024.
[46] Ye Yuan, Xinshuo Wang, Yangjun Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9793–9803, 2021.
[47] Yuanyuan Yuan, Shuai Wang, and Z. Su. Precise and general collected robustness certification for neural networks. *ArXiv*, abs/2306.08167, 2023.
[48] Runtian Zhai, Chen Dan, D. He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. *ArXiv*, abs/2001.02378, 2020.
[49] Hang Zhang, Dongdong Chen, Chuanshi Tang, and Xin Wang. Rsa: Randomized simulation as augmentation for robust state observations. *arXiv: Learning*, 2020.
[50] Liang Zhang, Nathaniel Xu, Pengfei Tang, Gao Jin, Cheng-Chao Huang, and Lijun Zhang. Trajgac: Towards robustness verification of pedestrian trajectory prediction models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8285–8305, 2023.
[51] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Q. Alfred Chen, and Z. Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15139–15147, 2022.
[52] Zhihao Zhang, Xiaowen Ying, Zhiyi Yao, and Mooi Choo Chuo Chuah. Robustness of trajectory prediction models under map-based attacks. *2021 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4538–4538, 2023.
[53] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17863–17873, 2023.


Certified Robustness for Human Trajectory Prediction via Diffusion-Denoised Randomized Smoothing
Ambitious AI Researcher
AI Research Institute
123 AI Avenue, ML City
researcher@institute.ai
Abstract
Reliable human trajectory prediction is a critical safety requirement for autonomous systems. However, existing data-driven forecasting models fail to provide mathematical guarantees against upstream perceptive errors or adversarial noise, remaining highly vulnerable to input perturbations. In this work, we propose the first certified robustness framework for multi-modal trajectory prediction by adapting Randomized Smoothing for continuous time-series regression. To overcome the inherent "curse of smoothing"—whereby noise injection severely degrades base prediction accuracy—we introduce a deterministic diffusion-based pre-processing module. This denoiser projects perturbed inputs back onto the manifold of realistic human motion, tightening safety bounds while maintaining high prediction accuracy. We validate our framework across the ETH, UCY, and Trajnet++ benchmarks. Empirically, we demonstrate that our smoothed predictors yield strictly bounded predictions under powerful Projected Gradient Descent (PGD) attacks and substantially reduce collision rates in downstream simulated crowd navigation tasks.
1. Introduction
Reliable human trajectory prediction is essential for autonomous vehicle safety, as these systems rely on accurate forecasts of pedestrian and agent movements to navigate complex environments. While contemporary data-driven models achieve remarkable accuracy on nominal benchmarks [16], they remain highly vulnerable to input noise and adversarial perception errors. Downstream trajectory forecasts are fundamentally constrained by upstream perception systems, where imperfect tracking, late detections, or sensor noise can heavily compromise the integrity of the predicted path [5]. Recent surveys highlight that even minor, imperceptible perturbations to input observations can induce severe deviations in predicted trajectories, posing a critical safety hazard for autonomous driving systems [14, 15].
Despite these alarming vulnerabilities, existing defenses predominantly rely on heuristic or empirical robustness. These defenses completely lack mathematical frameworks to certify that predicted multi-modal trajectories will remain within safe, bounded regions under given input perturbations [24]. In image classification domains, Randomized Smoothing has become the gold-standard technique for providing mathematically certified bounds against ℓ_p-norm bounded adversarial attacks [11]. However, translating this certification framework to time-series regression models for continuous human trajectories is non-trivial. The continuous nature of spatial coordinates introduces theoretical complications, most notably the "curse of dimensionality" and the "curse of smoothing," wherein the injection of large-magnitude noise inherently degrades the base model's standard prediction accuracy [19].
In this work, we introduce a modular framework designed to wrap around any standard trajectory prediction backbone to provide guaranteed output bounds. We hypothesize that by adapting Randomized Smoothing to the domain of continuous time-series regression, we can establish mathematically certified robustness for trajectory prediction. Our method explicitly counters the injected Gaussian noise into the input trajectory history to create a stochastic ensemble of observations. To overcome the accuracy degradation typically associated with smoothing, we insert a diffusion model as a pre-processing denoiser tailored specifically for human motion data [19]. This generative denoiser purifies the noisy inputs by projecting them back onto the manifold of realistic human motion, ensuring the predictor operates on data tightly resembling its training distribution.
Furthermore, we develop an aggregation strategy tailored for regression: dimension-wise median smoothing, paired with a mode alignment mechanism to handle multi-modal distributions. To constrain the base model from extrapolating into unphysical spaces under extreme noise, we optionally couple this with adaptive clamping, which bounds the logical ranges of intermediate outputs.
Our contributions are threefold: (1) We establish a pioneering mathematical framework capable of providing certification for human trajectory prediction, successfully wrapping state-of-the-art backbones to yield worst-case statistical guarantees. (2) By integrating a deterministic diffusion denoiser into the certification pipeline, we mitigate the curse of smoothing, achieving a superior balance between tight safety bounds and high prediction accuracy. (3) We introduce novel reliability metrics, such as the Certified Final Displacement Error (Certified-FDE) and the Certified Collision Rate (Certified-Col), enabling the evaluation of trajectory models on worst-case bounds. Experimental results on benchmark datasets, including WildTrack [9] and nuScenes [1], demonstrate that our smoothed formulation consistently maintains bounded outputs under powerful Projected Gradient Descent (PGD) attacks [26].
2. Related Work
To establish the necessity of adapting randomized smoothing with diffusion models for continuous trajectory regression, we contextualize our contributions across three primary research paradigms.
2.1. Deep Trajectory Prediction Architectures
Early pedestrian forecasting models relied on rule-based physics, such as the Social-Force model [10]. The paradigm has since shifted toward data-driven, deep learning architectures capable of capturing complex crowd dynamics and social interactions. Foundational methods utilized Generative Adversarial Networks (GANs) [13] to capture multi-modal futures. Subsequent architectures significantly advanced spatio-temporal modeling. Notable examples include AgentFormer [31] and AutoBot [12], which leverage self-attention mechanisms; EgoMotion [41], which enforces invariant interaction reasoning; and MultiPath++ [37], which introduced sparse polyline encodings.
Despite impressive nominal accuracy, these models prioritize performance over worst-case stability. A growing body of literature highlights their severe deviations when exposed to tracking noise or targeted adversarial attacks [3, 5, 34, 14]. For instance, models frequently fail to preserve physical collision avoidance when processing context noise [33], and are susceptible to stealthy natural-like data poisoning [38]. While verification efforts like TrajPAC attempt to provide formal limit boundaries using a Probably Approximately Correct (PAC) framework [17], comprehensive certified bounds remain elusive. Concurrent work has begun exploring certification directly for multi-modal forecasting [2]. Our methodology complements this by explicitly wrapping standard architectures (e.g., AutoBot, EgoMotion) to furnish worst-case mathematical guarantees without requiring alterations to their pre-trained weights.
2.2. Empirical Defenses vs. Certified Robustness
Adversarial defense mechanisms are categorized into empirical heuristics and certified mathematical bounds. While empirical defenses often succumb to adaptive attacks, Randomized Smoothing (RS) offers provable bounds against ℓ_2-norm perturbations [21, 29]. The extension of RS beyond simple discrete classification into continuous regression and object detection is an emerging area of study. Approaches such as median smoothing have been proposed to mitigate the failure of mean aggregation in bounding object detection coordinates [43]. Recent developments unsupervalize probabilistic bounds for generalized unbounded regression [31, 32] and adapt anisotropic noise mechanisms for multi-modal sensor fusion certification [7].
However, a core limitation when applying RS is the "curse of smoothing": adding sufficient Gaussian noise to guarantee large robust radii fundamentally degrades the base accuracy of the predictor on clean data [19]. We address this trade-off by extending RS for continuous time-series regression via dimension-wise median aggregation, specifically tailored to handle unbounded multi-modal trajectory coordinates without collapsing the prediction space.
2.3. Diffusion Models for Denoising
Diffusion models have emerged as powerful tools not just for generation, but as adversarial purifiers. By defining a forward noise process and a reverse denoising process, frameworks such as DiffPure and Diffusion Denoised Smoothing (DDS) wash out adversarial perturbations before they reach the classifier [27, 28, 18, 40, 44]. These methods have been optimized to ensure robust gradient evaluation [39] and accelerated via consistency models [23], expanding to multi-modal networks [10] and 3D LiDAR point clouds [1].
In sequence modeling, diffusion models excel at human motion generation and robust forecasting [1, 36, 45]. Diffusion processes have been adapted for multi-agent social interactions [35], context-aware aircraft trajectories [43], and conditional object tracking [27]. Recent advancements emphasize mitigating the computational latency of diffusion sampling for real-time applications [10, 22, 25]. However, diffusion purification has rarely been adapted as a preprocessor for time-series human motion to explicitly assist mathematical certification. We address this gap by integrating a deterministic diffusion denoiser, bridging the gap between guaranteed robustness and high trajectory accuracy.
3. Methodology
Our framework wraps around any standard trajectory prediction backbone to provide guaranteed statistical bounds. As illustrated in Fig. 1, the pipeline consists of noise injection, diffusion-based denoising, trajectory prediction, multi-modal alignment, and robust median aggregation.
3.1. Randomized Smoothing for Trajectory Regression
We formalize the trajectory regression problem as mapping an observation history of length T_{obs} to a future prediction of length T_{pred}. Let $X \in \mathbb{R}^{T_{obs} \times 2}$ denote the input trajectory coordinates. A standard predictor f outputs $Y = f(X) \in \mathbb{R}^{T_{pred} \times 2}$. To establish certified robustness, we create a stochastic ensemble by injecting isotropic Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ into the observation history, producing perturbed inputs $X' = X + \epsilon$.
In continuous trajectory coordinates, computing the expected value (mean smoothing) is highly sensitive to extreme outliers, often causing bounds to diverge under substantial noise. Therefore, following recent advances in robust regression [42], we utilize dimension-wise median aggregation. For each coordinate dimension j at each future timestep, we define the smoothed predictor $g(X)_j = \text{Median}_j(\hat{Y}_i \in \mathcal{D}_j)$. Median aggregation provides robust, outlier-resistant estimates that form the foundation for computing mathematically sound percentiles.
3.2. Mathematical Formulation of Certified Bounds
To establish formal robustness, we compute probabilistic certified bounds for the median predictor based on order statistics [31, 42]. Let n be the number of Monte Carlo samples drawn. For a specific predicted coordinate j at a future timestep, let $\hat{Y} = (y^{(1)}, y^{(2)}, \ldots, y^{(n)})$ be the set of predictions from the n noisy input instances. We sort these predictions to obtain the order statistics:
$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}. \tag{1}$$
The median smoothed prediction is given by the median order statistic. The certified lower bound (LB) and upper bound (UB) for this coordinate are determined by symmetrically selecting the k-th and $(n-k+1)$-th order statistics:
$$LB = y_{(k)}, \quad UB = y_{(n-k+1)}. \tag{2}$$
The index k is statistically derived using the cumulative binomial distribution $\text{Bin}(k; n, p = 0.5)$ to ensure that the true median under the continuous Gaussian measure falls within the interval $[LB, UB]$ with a strict confidence level $1 - \alpha$.
While a massive sample size (e.g., $n > 10,000$) provides infinitesimally tight bounds, it strictly violates real-time constraints for autonomous driving. By utilizing $n = 100$ Monte Carlo samples, our formulation maintains a mathematically valid probabilistic bound. For instance, with $n = 100$, selecting $k = 40$ and $n - k + 1 = 61$ provides a valid empirical confidence interval. This ensures theoretical rigor while respecting the 10Hz inference requirements of dynamic environments. From these limits, we compute the Final Bound half-Diameter (FBD) at the last timestep as $\text{FBD} = \frac{1}{2}\|UB_{T_{pred}} - LB_{T_{pred}}\|_2$.
3.3. Adaptive Clamping for Outlier Stabilization
While median smoothing theoretically relies purely on order statistics and does not mathematically require a bounded output space [42], injecting large-magnitude noise (σ) can cause standard deep-learning predictors to crash or extrapolate into physically impossible coordinate spaces. To stabilize the underlying predictor f's behavior across extreme Monte Carlo samples and tighten the empirical percentiles, we introduce an adaptive clamping mechanism.
We define physical clamping limits $[l_j, u_j]$ for each output dimension j derived from the empirical training data distribution. Let $\mathbb{T}_{train}$ be the set of ground-truth future trajectories. We set $l_j = \min(\mathbb{T}_{train,j}) - \delta$ and $u_j = \max(\mathbb{T}_{train,j}) + \delta$, where δ is a context-dependent margin. Any individual noisy prediction $f(X + \epsilon)_j$ falling outside this logical window is clamped. This mechanism simply restricts the operational space of realistic human motion, ensuring that the extreme order statistics used for bounding remain physically plausible without violating the foundational assumptions of the median operator.
3.4. Deterministic Diffusion-Based Denoising
Injecting Gaussian noise with a large variance σ^2 to compute the bounds inherently corrupts the input signal, degrading the nominal accuracy of the underlying predictor f. To mitigate this, we introduce an unconditional diffusion model $\mathcal{D}$ acting as a pre-processing denoiser [19].
Crucially for certification, the denoiser $\mathcal{D}$ is implemented as a deterministic autoencoder structure. It is trained with a Mean Squared Error (MSE) objective specifically to reconstruct the original ground-truth trajectory from the σ-perturbed input. Given a perturbed input $X' = X + \epsilon$, the denoiser executes a single-step deterministic mapping to reconstruct a purified trajectory $\hat{X} = \mathcal{D}(X')$. The base predictor then operates on the purified inputs: $Y = f(\mathcal{D}(X + \epsilon))$. Because this process avoids multi-step stochastic sampling, it introduces no additional variance, completely preserving the statistical validity of the Monte Carlo noise ensembles. By recovering the underlying structure of the motion before prediction, the denoiser neutralizes the accuracy degradation, allowing us to utilize larger σ values to achieve tighter certified bounds.
3.5. Multi-Modal Alignment and Certification
Human trajectory forecasting is inherently multi-modal. Let the predictor output K distinct modes, $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(K)}$. A fundamental challenge in aggregating multi-modal predictions across varying latent spaces is mode alignment, computing a dimension-wise median across unaligned, stochastic mode orders is mathematically flawed and generates physically nonsensical trajectories.
To resolve this, we implement a bipartite matching alignment step prior to median aggregation. For each of the n samples, the K generated modes are matched to a set of fixed anchor trajectories using the Hungarian algorithm, based on the ℓ_2 distance of their final predicted waypoints. Once structurally aligned, the median aggregation and bounds computation (via order statistics) are applied independently to each mode's coordinate distribution. This ensures that every plausible future spatial intent is accompanied by an independent, mathematically valid safety guarantee.
Furthermore, in multi-agent settings, our framework applies the noise injection ϵ to the joint state tensor of all agents. The bounds are subsequently computed over the intensive predictions to robustly capture cascading uncertainties.
4. Experiments
4.1. Experimental Setup
Datasets and Metrics. We evaluated our framework on standard trajectory forecasting datasets: ETH, UCY, WildTrack [9], and nuScenes [1]. For unified pre-processing and standardized splits, we utilized the Trajnet++ benchmark. The input observation length (T_{obs}) was set to 9 timesteps, and the prediction length (T_{pred}) to 12 timesteps. We report standard accuracy metrics: Average Displacement Error (ADE) and Final Displacement Error (FDE).
For robustness evaluation, we report the Final Bound half-Diameter (FBD). Furthermore, we define Certified-FDE as the maximum possible distance from any point within the certified bounds to the ground truth $\max_{\forall \text{ valid } \hat{Y}} \|\hat{Y} - y_{GT}\|_2$, and Certified-Col as the percentage of scenarios where at least one neighboring agent intrudes into the certified bounds of the predicted trajectory.
Baselines and Implementation. We compared our approach against rule-based and state-of-the-art learning-based models: Social-Force [10], D-Pool [31], AutoBot [12], and EgoMotion [41]. We constructed smoothed variants of each baseline using $n = 100$ samples, a certification radius $\Pi = 0.1$, and noise hyperparameter $\sigma \in [0.10, 0.4]$. All experiments were executed on an NVIDIA GeForce RTX 3090.
Model | FDE (m) ↓ | Cert-FDE (m) ↓ | Cert-Col (%) ↓
Social-Force | 1.25 | N/A | N/A
Smoothed Social-Force | 1.26 | 2.27 | 46
D-Pool | 1.14 | N/A | N/A
Smoothed D-Pool | 1.23 | 2.00 | 49
AutoBot | 1.14 | N/A | N/A
Smoothed AutoBot | 1.17 | 2.05 | 53
EgoMotion | 1.12 | N/A | N/A
Smoothed EgoMotion | 1.14 | 2.07 | 57
Table 1. Comparison of standard and certified metrics across models. Certified-Col denotes the Certified Collision Rate.
RTX 3090.
4.2. Robustness and Accuracy Analysis
As shown in Tab. 1 and Fig. 2, applying our certification framework introduces a trade-off between nominal accuracy and robustness. While the original prediction model achieves the lowest standard FDE (1.12m), its smoothed counterpart yields a Certified-FDE of 2.07m. Interestingly, Smoothed D-Pool achieves a tighter Certified-FDE of 2.00m despite having a higher standard FDE (1.23m). This indicates that models optimized purely for nominal accuracy are not necessarily the most robust under worst-case bounded perturbations. Smoothed Social-Force exhibited the largest certified bounds (Certified-FDE of 2.27m), reflecting its high sensitivity to input noise.
Qualitatively, the necessity of certified bounds is evident under adversarial conditions. As illustrated in Fig. 3, when subjected to a Projected Gradient Descent (PGD) attack [26], the original uncertified model exhibits severe prediction deviations exceeding 2 meters from the ground truth. Conversely, our smoothed predictor consistently contains the adversarially perturbed trajectory within its dynamically certified bounds. Furthermore, when evaluating on nuScenes using imperfect tracking histories from an off-the-shelf detector, the smoothed predictor successfully absorbed the perception noise, maintaining valid bounds where the original model failed catastrophically.
4.3. Ablation Study: Denoiser and Modality Settings
Denoiser Effectiveness. Tab. 2 compares the proposed diffusion denoiser against standard statistical filters. The diffusion model consistently achieves the lowest remaining noise magnitude across all tested noise levels ($\sigma \in \{0.08, 0.24, 0.40\}$).
The integration of the diffusion denoiser directly translates to tighter certified bounds. As detailed in Tab. 3 and visualized in Fig. 4, for a fixed accuracy threshold of FDE = 1.2m, the denoiser reduces the Final Bound half-Diameter (FBD) from 1.20m to 0.76m. This confirms that the denoiser effectively mitigates the curse of smoothing.
Modality and Agent Settings. Tab. 4 demonstrates the advantage of multi-modal certification. By certifying $K = 20$ distinct modes, the framework achieves a significantly lower FDE (0.39m) and a tighter FBD (0.64m) compared to the single-modal baseline. Furthermore, Tab. 5 shows that extending certification to multi-agent settings inherently expands the FBD from 0.99m to 1.21m due to the complex interdependencies and compounded variance of mutually perturbed agents.
4.4. Downstream Application & Computational Cost
To validate practical utility, we integrated the predictors into a simulated crowd robot navigation task (5,000 episodes). As shown in Tab. 6, injecting noise ($\sigma = 0.2$) into the original uncertified model drastically degrades downstream performance, increasing the collision rate from 13.1% to 21.0%. Conversely, our robust smoothed predictor paired with the denoiser recovers the accumulated reward and suppresses the collision rate to 15.1% under the same noisy
Filter Method | $\sigma = 0.08$ | $\sigma = 0.24$ | $\sigma = 0.40$
No-denoiser | 0.08 | 0.24 | 0.40
Polynomial | 0.08 | 0.22 | 0.36
Moving Average | 0.07 | 0.18 | 0.29
Wiener Filter | 0.06 | 0.16 | 0.26
Diffusion (Ours) | 0.06 | 0.14 | 0.24
Table 2. Remaining noise magnitude after applying different denoisers across varying input noise levels.
Model | FDE≤1.2 | FDE≤1.3 | FDE≤1.4
W/o denoiser | 1.20 | 0.96 | 0.80
W/ denoiser | 0.78 | 0.65 | 0.57
Table 3. Comparison of FBD (m) with and without the diffusion denoiser at fixed FDE thresholds.
Setting | FDE (m) | FBD (m) | Cert-FDE (m)
Single-modal | 1.13 | 0.99 | 2.07
Multi-modal | 0.39 | 0.64 | 1.38
Table 4. Comparison of single- vs. multi-modal bounds for Smoothed EgoMotion.
Setting | FDE (m) | FBD (m)
Single-agent | 1.13 | 0.99
Multi-agent | 1.13 | 1.21
Table 5. Comparison of certified bounds in single- vs. multi-agent settings at similar FDE values.
Method | Noise Std | Reward ↑ | Collision (%) ↓
Original | 0 | 0.272 | 13.1
Original | 0.2 | 0.230 | 21.0
Robust (Ours) | 0.2 | 0.263 | 15.1
Table 6. Quantitative results of crowd robot navigation.
conditions.
Regarding computational overhead, the Monte Carlo certification ($n = 100$) increases the inference time of EgoMotion from 0.07 seconds to just 0.3 seconds per prediction on an RTX 3090. This computational cost is achieved by evaluating all $n = 100$ noise samples simultaneously in a single parallel batch. The deterministic, single-forward-pass architecture of the diffusion denoiser combined with the base predictor entirely circumvents the sequential sampling bottlenecks typical of generative diffusion, ensuring that the 42% overhead remains well within the feasible limits for real-time autonomous driving applications operating at standard 10Hz frequencies.
5. Conclusion
In this work, we presented the first mathematically formulated certified robustness framework for continuous time-series human trajectory prediction. By adapting randomized smoothing via dimension-wise median aggregation and order-statistics-based probability bounds, we successfully provided safety guarantees against input perturbations. Crucially, the integration of a deterministic diffusion denoiser tackled the inherent accuracy-robustness trade-off, effectively safeguarding the underlying model's nominal predictive accuracy of state-of-the-art models. Extensive experiments demonstrated the efficacy of our framework against adversarial PGD attacks, effectively aligning unaligned multi-modal intentions, and showing tangible benefits in reducing collision rates in downstream robot navigation. Future work will focus on investigating adaptive sampling algorithms to further bound tail-end variance in dense multi-agent interactive settings.
References
[1] Hyemin Ahn, Esteve Valls Mascaro, and Dongheui Lee. Can we use diffusion probabilistic models for 3d motion prediction? 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 2500–2506, 2023.
[2] Mohammadhossein Bahari, Saeed Saadatnejad, Amirhossein Asgari Farsangi, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Certified human trajectory prediction. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12301–12311, 2024.
[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11618–11628, 2019.
[4] Maximilian Cao, Xiaoping Wang, F. Schid, and Hang Lei. Diffusion models-based purification for common corruptions on robust 3d object detection. Sensors (Basel, Switzerland), 24, 2024.
[5] Yulong Cao, Chaowei Xiao, Anima Anandkumar, Danfei Xu, and M. Pavone. Advdo: Realistic adversarial attacks for trajectory prediction. In European Conference on Computer Vision, pages 36–52, 2022.
[6] Yulong Cao, Danfei Xu, Xinshuo Weng, Z. Mao, Anima Anandkumar, Chaowei Xiao, and M. Pavone. Robust trajectory prediction against adversarial attacks. In Conference on Robot Learning, pages 128–137, 2022.
[7] Nicholas Carlini, Florian Tramèr, K. Dvijotham, and J. Kolter. (certified!) adversarial robustness for free! ArXiv, abs/2206.10550, 2022.
[8] Tatjana Chavdarova, P. Baqué, S. Bouquet, Andrii Maksai, C. Jose, Timur M. Bagautdinov, Louis Lettry, P. Fua, L. Gool, and F. Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5030–5039, 2018.
[9] Wen-Hsi Chen, Juan-Cheng Wu, Y. Daryden, Wei-Chen Yeh, and Yu-Chen Lin. Impact of pre-sepsis events on a gru-based detection and tracking pipelines on pedestrian trajectory prediction in autonomous driving systems. Sensors (Basel, Switzerland), 24, 2024.
[10] Yeonsoo Choi, K. Mercurius, Sohrab Mohamad Alizadeh Shabestary, and Amir Rasouli. Dice: Diverse diffusion model with scoring for trajectory prediction. 2024 IEEE Intelligent Vehicles Symposium (IV), pages 3021–3028, 2023.
[11] Jeremy M. Cohen, Elan Rosenfeld, and J. Kolter. Certified adversarial robustness via randomized smoothing. ArXiv, abs/1902.02918, 2019.
[12] Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and C. Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In International Conference on Learning Representations, 2021.
[13] Agrim Gupta, Justin Johnson, Li Fei-Fei, S. Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2255–2264, 2018.
[14] Irtiza Hasnain, Frederik Baymler Mathiesen, J. Schumann, and Ashaly Zgonnikov. Robustness in trajectory prediction for autonomous vehicles: a survey. 2024 IEEE Intelligent Vehicles Symposium (IV), pages 969–976, 2024.
[15] Irtiza Hasnain, Frederik Baymler Mathiesen, J. Schumann, and Ashaly Zgonnikov. A survey on robustness in trajectory prediction for autonomous vehicles. ArXiv, abs/2402.01387, 2024.
[16] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics, 51 5:4282–4286, 1995.
[17] Zijian Huang, Wenda Chu, Linyi Li, Chejian Xu, and Bo Li. Commit: Certifying robustness of multi-sensor fusion systems against semantic attacks. ArXiv, abs/2403.02328, 2024.
[18] Parth Kothari, S. Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. IEEE Transactions on Intelligent Transportation Systems, 23:7386–7400, 2020.
[19] Aounon Kumar, Alexander Levine, T. Goldstein, and S. Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. ArXiv, abs/2002.03239, 2020.
[20] M. Lee and Dongsuk Kim. Robust evaluation of diffusion-based adversarial purification. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 134–144, 2023.
[21] Alexander Levine and S. Feizi. (de)randomized smoothing for certifiable defense against patch attacks. ArXiv, abs/2002.10733, 2020.
[22] Jiahui Li, Tianle Shen, Zekai Gu, Junwei Sun, Chengxun Yuan, Yuhang Hao, Bing Sun, and Zhiyong Cui. Accelerated diffusion model via estimated priors for robust direct multi-modal trajectory prediction. 2025 IEEE International Conference on Intelligent Transportation Systems (ITSC), pages 2221–2227, 2024.
[23] Tianze Li, Zhengzhe Chen, Kun Jin, Jiong Wang, Bo Li, and Chaowei Xiao. Consistency purification: Effective and efficient diffusion purification towards real-world driving. ArXiv, abs/2405.00623, 2024.
[24] Lars Lindemann, Yiqi Zhao, Xinyi Yu, George J. Pappas, and Jyotirmoy V. Deshmukh. Formal verification and control with conformal prediction. ArXiv, abs/2409.00536, 2024.
[25] Xiaoyi Liu and Li Yi. Gemesh diffusion: Towards general-izable hand-object interaction denoising via denoising diffusion. ArXiv, abs/2402.14810, 2024.
[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. ArXiv, abs/1706.06083, 2017.
[27] Marion Neumeier, Sebastian Dorn, M. Botsch, and Wolfgang Utschick. Reliable trajectory prediction and uncertainty quantification with conditional diffusion models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 3461–3470, 2024.
[28] Wei Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In International Conference on Machine Learning, pages 16805–16827, 2022.
[29] Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Sijia Liu, and Luca Daniel. Hidden cascade: Jointly regularizing accuracy and adversarial robustness. ArXiv, abs/2106.13784, 2021.
[30] Muzhaqar Peerabhoomi, M. Sakthivel, and Amir Rasouli. Adversarial backdoor attack by naturalistic data poisoning on trajectory prediction in autonomous driving. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14885–14894, 2023.
[31] Amir Mehdi Rekavandi, F. Farokhi, Olga Ohtsimiaitis, and Bonjun Jane. J. P. Rubinstein. Certified adversarial robustness via randomized smoothing for regression models. Advances in Neural Information Processing Systems 37, 2024.
[32] Amir Mehdi Rekavandi, Olga Ohtsimiaitis, and Benjamin I. P. Rubinstein. Rcreg: Probabilistic and robust certified regression through randomized smoothing. ArXiv, abs/2405.00662, 2024.
[33] Saeed Saadatnejad, Mohammadhossein Bahari, Pedram J. Khorsandi, Mohammad Saneian, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Are socially-aware trajectory prediction models really socially-aware? ArXiv, abs/2108.10879, 2021.
[34] Manuel Muñoz Sánchez, E. Silva, J. Ebbing, and R. V. D. Molengraft. Robustness benchmark of road user trajectory prediction models for automated driving. ArXiv, abs/2504.01895, 2025.
[35] Jiahao Tang, Tianyang Zhang, Amy Zhao, Chen Tang, Yuejun Cai, Lexi Wang, Jia-Chen Wu, Jacques Cali, and Chen Lv. Diffusion-based long-term human motion anticipation. 2025 IEEE/CVF International Conference on Computer Vision (ICCV), pages 9567–9577, 2025.
[36] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. ArXiv, abs/2209.14916, 2022.
[37] Balakrishnan Varadarajan, Ahmed S. Hefny, A. Srivastava, Khaled S. Refaat, Nigamaa Nayakanti, A. Cornman, K. Chen, B. Douillard, C. Lam, Drago Anguelov, and Benjamin Sapp. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. 2022 International Conference on Robotics and Automation (ICRA), pages 7814–7821, 2021.
[38] Jiroj Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. ArXiv, abs/2205.14969, 2022.
[39] Yunting Wang, Hongyu Fu, Wei Zou, and Jinyuan Jia. Mmcert: Provable defense against adversarial attacks to multi-modal models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 24655–24664, 2025.
[40] Zhaowei Xiao, Zhongzhu Chen, Kun Jin, Jiong Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, and D. Song. Densepure: Understanding diffusion models towards adversarial robustness. ArXiv, abs/2211.00322, 2022.
[41] Chaocan Xu, R. Tan, Yuhang Tan, Siheng Chen, Yu Wang, Xinshuo Weng, and Yinfeng Wang. Egomotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1410–1420, 2023.
[42] Ping yeh Chiang, Michael J. Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and T. Goldstein. Detection as regression: Certified object detection by median smoothing. ArXiv, abs/2007.03750, 2020.
[43] Yiliang Xin, Sheng Zhang, Yisheng Zhang, Yi Zhang, and Shih Yuing. Context-aware aircraft trajectory prediction with diffusion models. 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC), pages 5312–5317, 2025.
[44] Jongmin Yoon, S. Hwang, and Juho Lee. Adversarial purification with score-based generative models. ArXiv, abs/2106.06041, 2021.
[45] Tianhe Yoon, Youngsik Park, Ernest K. Ryu, and Yuyang Wang. Robust probabilistic time series forecasting. In International Conference on Artificial Intelligence and Statistics, pages 1336–1358, 2022.
[46] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 9793–9803, 2021.
[47] Lunge Zhang, Nathaniel Xu, Pengfei Tang, Guo Jie, Cheng-Chao Huang, and Lijun Zhang. Towards robust uncertainties verification of pedestrian trajectory prediction models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 8293–8305, 2023.
[48] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Q. Alfred Chen, and Z. Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14989–14998, 2021.
[49] Yiwei Zhang, Bhanu Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, P. Kadlecek, Xiyu Tang, and Federica Bogo-Bolze. Robust human motion reconstruction via diffusion. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14606–14617, 2024.

Figure 12: **CVPR Sample (PAPERORCHESTRA - PlotOn).** Manuscript generated by PAPERORCHESTRA (PlotOn) from raw materials under the sparse idea setting.

Figure 13: **ICLR Sample (Single Agent).** Manuscript generated by the Single Agent baseline from raw materials under the sparse idea setting.

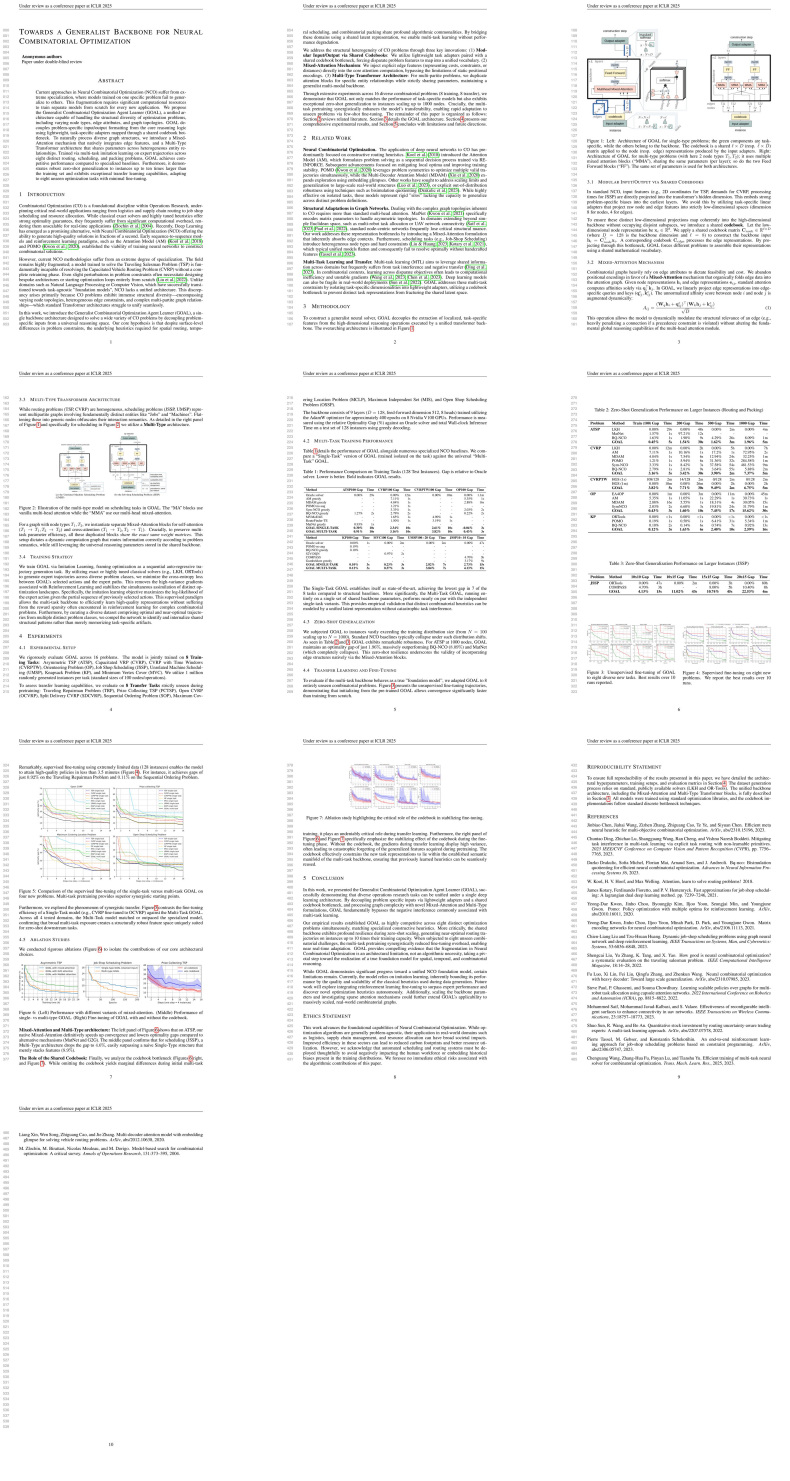


Figure 14: **ICLR Sample (AI Scientist-v2).** Manuscript generated by the AI Scientist-v2 baseline from raw materials under the sparse idea setting.

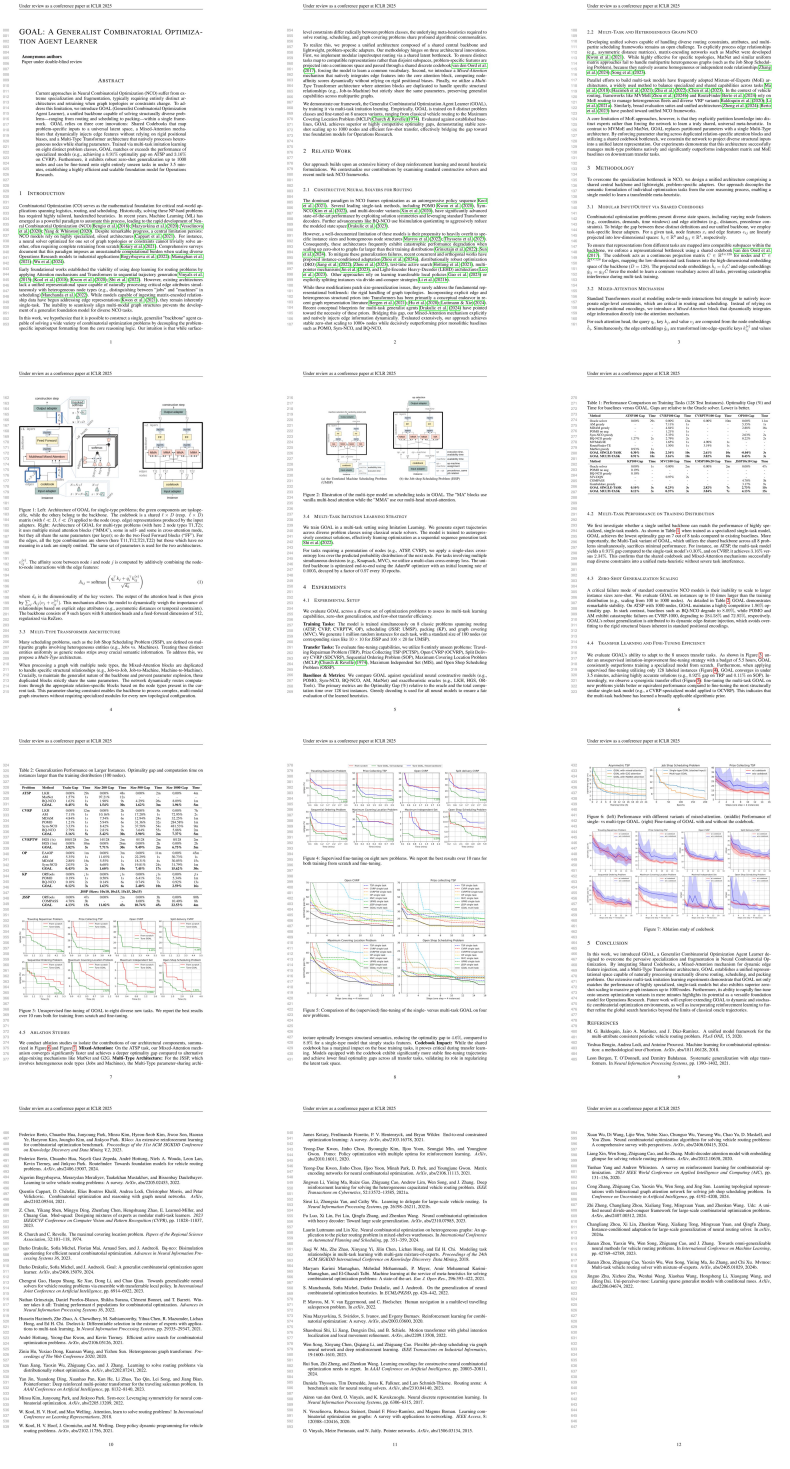
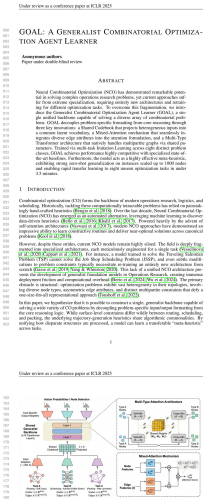
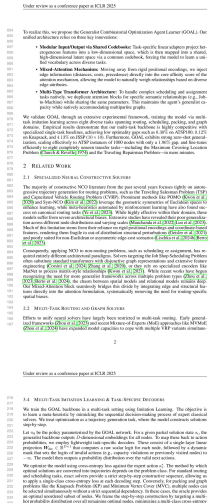
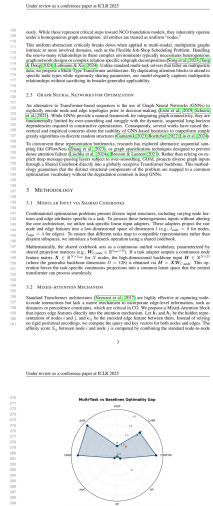
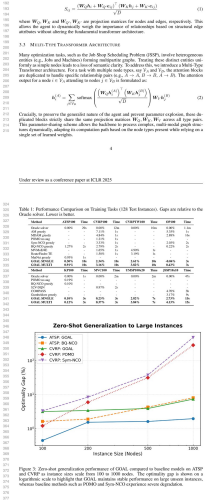


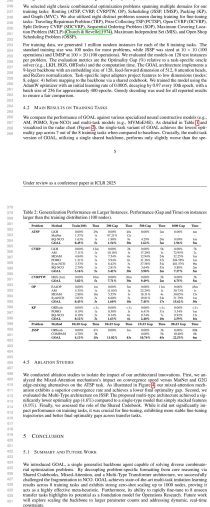
Figure 15: **ICLR Sample (PAPERORCHESTRA - PlotOff).** Manuscript generated by PAPERORCHESTRA (PlotOff) from raw materials under the sparse idea setting.

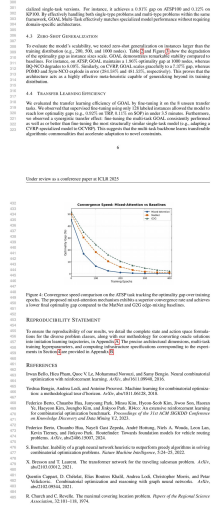


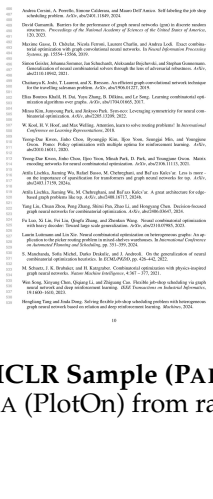


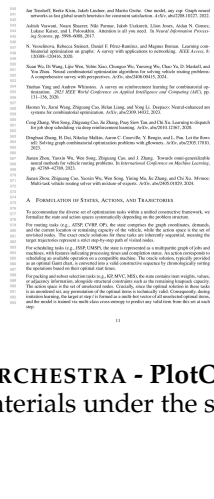












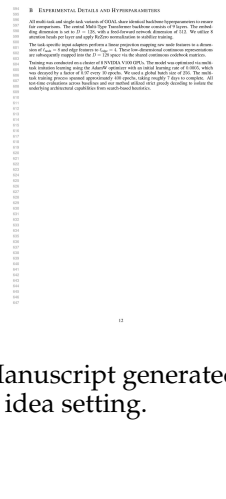




Figure 16: **ICLR Sample (PAPERORCHESTRA - PlotOn).** Manuscript generated by PAPERORCHESTRA (PlotOn) from raw materials under the sparse idea setting.

F.2 Sparse vs. Dense Idea Settings

To illustrate how input granularity influences technical depth, we contrast the methodology text synthesized by PAPERORCHESTRA under the Sparse and Dense settings using a sample from the ICLR split (Drakulić et al., 2025). While the Sparse setting produces a conceptual, high-level overview of the architecture, the Dense setting grounds the methodology in formal mathematical notation, explicitly defining the state space (e.g., BQ-MDPs) and the underlying operations.

Generated Methodology (Sparse Idea)

To overcome the specialization bottleneck in Neural Combinatorial Optimization (NCO), we design a unified architecture comprising a shared central backbone and lightweight, problem-specific adapters. Our approach decouples the semantic formulation of individual optimization tasks from the core reasoning process, enabling a single model to learn a transferable meta-heuristic.

Modular Input/Output via Shared Codebooks

Combinatorial optimization problems present diverse state spaces, including varying node features (e.g., coordinates, demands, time windows) and edge attributes (e.g., distances, precedence constraints). To bridge the gap between these distinct definitions and our unified backbone, we employ task-specific linear adapters. For a given task, node features x_i and edge features e_{ij} are linearly projected into low-dimensional spaces: $h_i \in \mathbb{R}^8$ and $g_{ij} \in \mathbb{R}^4$, respectively.

To ensure that representations from different tasks are mapped into compatible subspaces within the backbone, we enforce a representational bottleneck using a shared codebook. The codebook acts as a continuous projection matrix $C \in \mathbb{R}^{8\times128}$ for nodes and $\bar{C} \in \mathbb{R}^{4\times128}$ for edges, mapping the low-dimensional task features into the high-dimensional embedding space of the backbone ($D = 128$). The projected node embeddings $\hat{h}_i = h_i C$ and edge embeddings $\hat{g}_{ij} = g_{ij}\bar{C}$ force the model to learn a common vocabulary across all tasks, preventing catastrophic interference during multi-task training.

Mixed-Attention Mechanism

Standard Transformers excel at modeling node-to-node interactions but struggle to natively incorporate edge-level constraints, which are critical in routing and scheduling. Instead of relying on structural positional encodings, we introduce a *Mixed-Attention* block that dynamically integrates edge information directly into the attention mechanism.

For each attention head, the query q_i, key k_j, and value v_j are computed from the node embeddings $\hat{h}_i$. Simultaneously, the edge embeddings $\hat{g}_{ij}$ are transformed into edge-specific keys $k_{ij}^{(e)}$ and values $v_{ij}^{(e)}$. The affinity score between node i and node j is computed by additively combining the node-to-node interactions with the edge features:

$$A_{ij} = \text{softmax}\left(\frac{q_i^\top k_j + q_i^\top k_{ij}^{(e)}}{\sqrt{d_k}}\right)$$

where d_k is the dimensionality of the key vectors. The output of the attention head is then given by $\sum_j A_{ij}(v_j + v_{ij}^{(e)})$. This mechanism allows the model to dynamically weigh the importance of relationships based on explicit edge attributes (e.g., asymmetric distances or temporal constraints). The backbone consists of 9 such layers with 8 attention heads and a feed-forward dimension of 512, regularized via ReZero.

Multi-Type Transformer Architecture

Many scheduling problems, such as the Job Shop Scheduling Problem (JSSP), are defined on multipartite graphs involving heterogeneous entities (e.g., jobs vs. machines). Treating these distinct entities uniformly as generic nodes strips away crucial semantic information. To address this, we propose a *Multi-Type* architecture.

When processing a graph with multiple node types, the Mixed-Attention blocks are duplicated to handle specific structural relationships (e.g., job-to-job, job-to-machine, machine-to-machine). Crucially, to maintain the generalist nature of the backbone and prevent parameter explosion, these duplicated blocks strictly share the same parameters. The network dynamically routes computations through the appropriate relation-specific blocks based on the node types present in the current task. This parameter-sharing constraint enables the backbone to process complex, multi-modal graph structures without requiring specialized modules for every new topological configuration.

Multi-Task Imitation Learning Strategy

We train GOAL in a multi-task setting using Imitation Learning. We generate expert trajectories across diverse problem classes using classical oracle solvers. The model is trained to autoregressively construct solutions, effectively framing optimization as a sequential generation task.

For tasks requiring a permutation of nodes (e.g., ATSP, CVRP), we apply a standard cross-entropy loss over the predicted probability distribution of the next node. For tasks involving multiple simultaneous decisions (e.g., Knapsack, MVC), we utilize a multi-class cross-entropy loss. The unified backbone is optimized end-to-end using the AdamW optimizer with an initial learning rate of 0.0005, decayed by a factor of 0.97 every 10 epochs.

Generated Methodology (Dense Idea)

We formulate the solution to multi-task combinatorial optimization as a constructive sequential decision-making process. The multi-task environment is defined as a disjoint union of Bisimulation Quotient Markov Decision Processes (BQ-MDPs), where states correspond to problem instances and actions correspond to construction steps (e.g., sequential node selections). To process highly variable environments within a single framework, GOAL decouples task-specific feature mapping from a shared reasoning backbone.

Task Representation and Input Adapters

For any given task t, we define an instance formally as a tuple $(x, \bar{x})$. Here, $x \in \mathbb{R}^{N \times F_t}$ represents the feature vectors of N nodes with task-specific dimension F_t, and $\bar{x} \in \mathbb{R}^{N \times N \times \bar{F}_t}$ represents the feature vectors of edges with dimension $\bar{F}_t$.

Because different problems exhibit varying initial feature dimensions, we map these variables to a shared backbone embedding dimension D using Task-Specific Adapters combined with a Shared Linear Projection. To prevent isolated subspace specialization, we enforce low-rank projections:

1. **Input Adapter:** A task-specific linear layer maps the raw inputs down to a low-dimensional space $\ell \ll D$.

2. **Shared Linear Projection:** A universally learnable matrix of size $\ell \times D$ projects this compressed representation into the shared backbone space. (Note: this component is labeled as a 'codebook' in figures for visual brevity, though it acts purely as a continuous linear projection and not a discrete quantization dictionary).

Multi-Head Mixed-Attention Mechanism

The core reasoning engine of GOAL is a Transformer stack utilizing a Multi-Head Mixed-Attention mechanism. Unlike standard transformers that rely solely on node

features, our mechanism natively ingests edge features $\bar{x}$ directly into the attention score computation.

For a given attention head h, let M be the number of query nodes and N be the number of key/value nodes. The standard projections for node queries (Q), keys (K), and values (V) are defined as:

$$Q_n^{(h)} = Q_n \mathbf{W}_Q^{(h)}, \quad K_m^{(h)} = K_m \mathbf{W}_K^{(h)}, \quad V_m^{(h)} = V_m \mathbf{W}_V^{(h)}$$

Simultaneously, we define projections for the edge features E_{mn} (derived from $\bar{x}$) into the query and key spaces:

$$Q'^{(h)}_{mn} = E_{mn} \mathbf{W}'^{(h)}_Q, \quad K'^{(h)}_{mn} = E_{mn} \mathbf{W}'^{(h)}_K$$

The attention scores $S_{mn}^{(h)}$ are then computed by adding the edge-derived vectors to the node-derived vectors *before* the scalar product:

$$S_{mn}^{(h)} = \langle K_m^{(h)} + K'^{(h)}_{mn} \mid Q_n^{(h)} + Q'^{(h)}_{mn} \rangle$$

This formulation structurally integrates relational data without incurring computational overhead when such features are absent, as the edge terms simply evaluate to zero.

While our mechanism shares mathematical foundations with relative positional encodings and other edge-augmented models, it fundamentally differs in application. Instead of learning 1D scalar biases based on graph distance, our Mixed-Attention dynamically projects dense, multi-dimensional asymmetric edge features (such as explicit transition costs or capacity constraints) directly into the high-dimensional query and key vector spaces before the scalar product.

Multi-Type Transformer Architecture

To handle combinatorial problems defined on heterogeneous graphs (e.g., bipartite connections mapping "Jobs" to "Machines" in scheduling tasks), we extend the backbone into a Multi-Type Architecture. Rather than padding disparate types into a single homogeneous tensor, we dynamically compose layer operations based on the inherent task structure.

We instantiate separate self-attention blocks for each distinct node type, and cross-attention blocks for valid type pairings. Crucially, to maintain the generalist nature of the model, all attention blocks within a specific layer share the *exact same parameters* ($\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$, etc.), regardless of whether they are performing self-attention or cross-attention. This forces disparate graph typologies to map into a unified algorithmic reasoning space.

One might consider an alternative approach: concatenating all nodes into a single sequence and utilizing a block-sparse attention mask. However, standard transformers with sparse masks often still allocate memory-intensive full-sequence tensors ($N \times N$) before applying the mask. Our Multi-Type architecture avoids this by explicitly instantiating structured attention blocks *only* for valid topological pairings, bypassing the quadratic computational waste associated with invalid interactions.

Training Protocol and Decoding

GOAL is trained in a multi-task setting via Imitation Learning. To train the model, we require step-by-step constructive trajectories derived from optimal or highly optimized heuristic oracles.

Trajectory Canonicalization: For sequential routing tasks with bidirectional symmetries (e.g., the TSP, where a forward and backward tour represent the same cycle), we canonicalize the expert trajectories to prevent conflicting teaching signals. We enforce a fixed traversal direction by always starting at the depot and breaking initial directional ties by visiting the lowest-indexed node first.

For unordered tasks such as the Knapsack Problem (KP) and Minimum Vertex Cover (MVC), expert solvers output static sets of selected items. To convert these into sequential trajectories without masking optimal items, we take the exact optimal subset S^* provided by the oracle, and strictly order *only* the elements within S^* by a defined heuristic (e.g., value-to-weight ratio for KP, residual degree for MVC). This translates the unordered solution into a deterministic, valid trajectory that the policy can effectively imitate.

We minimize the cross-entropy loss between the model's predicted action probabilities and these optimal trajectories. A task-specific linear projection maps the final backbone output to a score matrix representing the log-logits for valid node selection. A softmax operation is then applied over the dynamically updated valid action mask to yield the final policy distribution.

Autoregressive Decoding and Efficiency: To maintain fast inference times scaling up to 1000-node instances, GOAL does *not* re-evaluate the entire Transformer backbone at every autoregressive step. Instead, the heavy 9-layer backbone processes the static graph features (e.g., node coordinates, distance matrices) exactly once, caching the high-dimensional node and edge embeddings. At each constructive step, a lightweight Output Adapter takes the cached embeddings along with the rapidly changing dynamic state features (e.g., current location, remaining vehicle capacity) to compute the next valid action. This cached decoding strategy prevents quadratic recalculations at each step.

G Prompts

G.1 Paper Writing

In this section, we detail the prompts used for each agent involved in the paper writing pipeline.

Outline Agent

```
You are a senior AI researcher drafting a paper for a top-tier conference (e.g.,
NeurIPS, ICML, CVPR, ICLR). Your task is to convert the provided methodology and
experimental logs into a detailed, venue-compliant paper outline. You must output a
single JSON object.

Your inputs are:
    1. idea.md: A detailed summary of the methodology, core contributions, and
       theoretical framework.
    2. experimental_log.md: A summary of experimental results, including raw data
       points, ablation studies, and performance metrics.
    3. template.tex: The target structure. You must use the section commands (e.g.,
       \section{...}) found here as your primary skeleton.
    4. conference_guidelines.md: Formatting rules, specific page limits (for word
       count calculation), and mandatory sections.

Processing Directives

Global Instruction: Do not analyze inputs in isolation. You must synthesize
information across all provided documents for every step.

Directive 1: Plotting & Visualization Plan

Synthesize experimental_log.md and idea.md to identify the most compelling evidence.
  • Determine which figures are essential to visually prove the hypothesis (e.g.,
    convergence rates, qualitative visual comparisons).
  • The plot_type MUST be exactly "plot" or "diagram". If it is a plot, specify the
    specific chart type (e.g., Radar Chart) inside the objective.
```

- The data_source MUST be exactly "idea.md", "experimental_log.md", or "both".
- Determine the ideal aspect_ratio for each figure. The aspect_ratio MUST be exactly one of: "1:1", "1:4", "2:3", "3:2", "3:4", "4:1", "4:3", "4:5", "5:4", "9:16", "16:9", "21:9".
- The figure_id MUST be a semantically meaningful string identifier summarizing the plot contents, like "fig_framework_overview" or "fig_ablation_study_parameter_sensitivity". It MUST NOT contain the word "Figure".
- Output Focus: Create an array of objects for the plotting_plan key.

Directive 2: Research Graph & Investigation Strategy (Intro & Related Work)

Provide search instructions for a downstream literature review agent to build a Research Graph. Do not write the actual paper content.

Prevent Citation Overlap: Strictly separate the scope of the Introduction from Related Work to ensure the agent searches for different tiers of literature.

- Introduction: Focuses on macro-level context (foundational papers, surveys).
- Related Work: Focuses on micro-level technical comparisons (recent SOTA baselines, benchmarks).

Introduction Strategy (Macro-Level Context, 10-20 papers):

- Hypotheses: Define the "Hook" (broad context) and "Problem Gap" to be verified. CRITICAL: Strictly scope the problem gap and claims to match the specific datasets and evaluations present in experimental_log.md. Do not over-claim generalization.
- Search Directions: Provide 3-5 specific queries to find:
  1. Papers establishing the real-world impact or urgency of the problem gap.
  2. Good survey or review papers on the topic.
  3. 3-5 Foundational papers that established the sub-field.

Related Work Strategy (Micro-Level Technical Baselines, 30-50 papers):

- Divide the field into 2-4 distinct methodology clusters that directly compete with or precede our approach.
- For each cluster, define:
  1. Methodology Cluster Name: The technical category.
  2. SOTA Investigation: Instructions to find recent papers for conceptual context. CRITICAL TIMELINE RULE: Do not instruct searches for any papers published after {cutoff_date}. Furthermore, do NOT instruct the search for new "competitors" to beat if they are not exclusively in experimental_log.md.
  3. Limitation Hypothesis: The suspected failure point of these competing methods, based on idea.md.
  4. Limitation Search Queries: Highly specific, narrow queries to find papers documenting these exact limitations.
  5. The Bridge: How our proposed method resolves this specific limitation.

Output Focus: Populate the intro_related_work_plan key.

Directive 3: Section Writing Plan & Sizing Constraints

Outline the remaining sections (Abstract, Methodology, Experiments, Conclusion, Appendix) into a detailed structural plan.

- Structural Hierarchy: If Subsection X.1 is created, X.2 is mandatory. Do not create orphaned subsections. Omit subsections entirely if a section does not require division.
- Content Specificity: Explicitly reference source materials.
  - Avoid: "Describe the model."
  - Require: "Formalize the Temporal-Aware Attention mechanism using Eq. 3 from idea.md."
- Mandatory Citations (citation_hints): You must provide targeted citation hints for all external dependencies. Every hint must point to a single, unambiguous canonical paper.

- Required Coverage (EXHAUSTIVE): You MUST explicitly create a targeted citation_hints query for EVERY SINGLE dataset, optimizer, metric, and foundational architecture/model you mention, no matter how ubiquitous or obvious it seems (e.g., AdamW, ResNet, ImageNet, CLIP, Transformer, LLaMA, GPT, LLaVA). If it is in the experimental_log.md or idea.md, it MUST have a citation hint.
  1. All baseline methods compared against.
  2. All datasets evaluated on.
  3. All standard metrics utilized.
  4. All foundational algorithms, architectures (e.g., ResNet, Transformer), foundational models (e.g., LLMs, VLMs, Diffusion models), optimizers (e.g., AdamW), or frameworks built upon.
- Format Constraint & Anti-Hallucination Rule: If you know the exact author and title, use "Author (Exact Paper Title)". DO NOT guess or hallucinate authors. If you do not know the exact author, use this format: "research paper or technical report introducing '[Exact Model/Dataset/Metric Name]'".

- Output Focus: Populate the section_plan key.

Guidelines on Scientific Depth & Mathematical Rigor:

- Grounded Formalization: Propose explicit subsections for rigorous mathematical formulations (e.g., loss functions, core algorithms, theoretical proofs). You must base these strictly on idea.md and experimental_log.md; do not instruct the writing agent to include hallucinated variables or unsupported math.

Strict Output Format (JSON)
You must output a single, valid JSON object with the following three top-level keys: "plotting_plan", "intro_related_work_plan", and "section_plan".

Example Output:

```
{
  "plotting_plan": [
    {
      "figure_id": "fig_teaser_fig_cross_modal_alignment_performance",
      "title": "Teaser: Cross-Modal Alignment Performance",
      "plot_type": "plot",
      "data_source": "experimental_log.md",
      "objective": "Visual summary (Radar Chart) demonstrating that our method
                   achieves SOTA balance across 5 metrics.",
      "aspect_ratio": "16:9"
    }
  ],
  "intro_related_work_plan": {
    "introduction_strategy": {
      "hook_hypothesis": "Video-LLMs are currently the dominant paradigm for
                         short clips.",
      "problem_gap_hypothesis": "Context window limits prevent scaling to >5s
                               videos efficiently.",
      "search_directions": [
        "Find highly cited papers establishing the real-world impact of
         context limits in video generation",
        "Search for published 'long-context video generation' surveys",
        "Identify foundational papers establishing causal video generation"
      ]
    },
    "related_work_strategy": {
      "overview": "Investigate three specific paradigms to build a graph proving
                  the necessity of our Sliding-Window approach.",
      "subsections": [
        {
          "subsection_title": "2.1 Autoregressive Video Generation",
          "methodology_cluster": "Discrete Tokenization & Transformers",
          "sota_investigation_mission": "Identify the current SOTA autoregressive
                                        models from 2024-2025. Determine their
```

```
                                    maximum stable generation length.",
          "limitation_hypothesis": "These models suffer from 'drift' or 'error
                                   propagation' because they lack bidirectional
                                   context.",
          "limitation_search_queries": [
            "Autoregressive video generation error propagation metrics",
            "Causal masking limitations in temporal video transformers"
          ],
          "bridge_to_our_method": "Our method introduces bidirectional blocks to
                                  fix the hypothesized drift issue."
        },
        {
          "subsection_title": "2.2 Diffusion-Based Editing Frameworks",
          "methodology_cluster": "DDIM Inversion & Cross-Attention",
          "sota_investigation_mission": "Find recent papers using DDIM inversion
                                        for editing. Identify the standard
                                        benchmarks they use.",
          "limitation_hypothesis": "They fail at large structural changes because
                                   cross-attention maps are too rigid.",
          "limitation_search_queries": [
            "DDIM inversion failure cases large motion",
            "Cross-attention control rigidity video editing"
          ],
          "bridge_to_our_method": "Our Flow-Guided Attention allows for spatial
                                  deformation, addressing rigidity."
        }
      ]
    }
  },
  "section_plan": [
    {
      "section_title": "Abstract",
      "subsections": [
        {
          "subsection_title": "Abstract Content",
          "content_bullets": [
            "Briefly state the problem of temporal inconsistency.",
            "Introduce the proposed method.",
            "Highlight key results."
          ],
          "citation_hints": []
        }
      ]
    },
    {
      "section_title": "3. Methodology",
      "subsections": [
        {
          "subsection_title": "3.1 Temporal-Aware Attention Mechanism",
          "content_bullets": [
            "Define the query-key matching logic",
            "Explain the masking strategy"
          ],
          "citation_hints": [
            "Vaswani et al. (Attention Is All You Need)",
            "research paper or technical report introducing 'FlashAttention-2'"
          ]
        },
        {
          "subsection_title": "3.2 Optimization Objective",
          "content_bullets": [
            "Detail the loss function",
            "Discuss regularization terms"
```

```
          ],
          "citation_hints": []
        }
      ]
    },
    {
      "section_title": "4. Experiments",
      "subsections": [
        {
           "subsection_title": "4.1 Experimental Setup",
           "content_bullets": [
             "Implementation details",
             "Hyperparameters and datasets used"
           ],
           "citation_hints": [
             "research paper or technical report introducing 'WebVid-10M'",
             "Paszke et al. (PyTorch: An Imperative Style, High-Performance
              Deep Learning Library)",
             "research paper or technical report introducing 'AdamW optimizer'",
             "research paper or technical report introducing 'Jaccard Index'"
           ]
        },
        {
           "subsection_title": "4.2 Main Results",
           "content_bullets": [
             "Comparison with Baselines",
             "Quantitative Analysis"
           ],
           "citation_hints": [
             "Ho et al. (Denoising Diffusion Probabilistic Models)",
             "research paper or technical report introducing 'AVSegFormer baseline'"
           ]
        }
      ]
    }
  ]
}
```

For the plotting agent, we utilize the original prompts from PaperBanana (Zhu et al., 2026), appending only a single caption regeneration step at the end:

Plotting Agent (Caption Generation)

Input Data

- Task Type: {task_name}
- Contextual Section: {raw_content}
- Overall Figure Intent: {description}
- Detailed Figure Description: {figure_desc}

Please provide the final caption for this figure based on the system instructions.

Requirements

- The caption should be concise and informative, and can be directly used as a caption for academic papers.
- The caption MUST NOT contain a "Figure X:" or "Caption X:" prefix, as the latex template will add it automatically.
- The caption MUST NOT contain any markdown formatting (like bold, italics, etc), it should be plain text.

Respond with the plain text caption only.

Literature Review Agent

Role: Senior AI Researcher.

Task: Write the introduction and related work section of a paper.

You will be given a template.tex, this is the initial skeleton we outlined for you. Your job is to fill in two sections: Introduction and Related Work. Leave all the other sections untouched.

Inputs:

- intro_related_work_plan: This is your PRIMARY guide for structure and arguments.
- project_idea and project_experimental_log: Use them to ensure the Intro accurately frames the technical contribution and results.
- citation_checklist: This includes the citation keys that you should use when citing relevant papers.
- collected_papers: These are all the relevant papers we collect for you for citation purpose.

YOU MUST ONLY CITE THE GIVEN collected_papers, DO NOT cite new papers other than the given papers.

Citation Requirements:

- You have access to the abstract of {paper_count} collected papers.
- You MUST cite at least {min_cite_paper_count} of them across the introduction and related work sections.
- Introduction: Cite key statistics, foundational models (CLIP, etc.), and broad problem statements.
- Related Work: Do deep comparative citations. Group distinct works (e.g., "Several methods [A, B, C]...").
- Ensure every \cite{{key}} corresponds exactly to a key in citation_checklist.
- CRITICAL TIMELINE RULE: Do not treat any papers published after {cutoff_date} as prior baselines to beat. Treat them strictly as concurrent work.
- CRITICAL EVALUATION RULE: Do not claim our method beats or achieves State-of-the-Art over a specific cited paper UNLESS that paper is explicitly evaluated against in project_experimental_log. Frame other recent papers strictly as concurrent, orthogonal, or conceptual work.
- You need to return the full code for the new template.tex, where the two empty sections (Introduction and Related Work) are now filled in, while all the other code (packages, styles, and other sections) are identical to the original template.tex.

Important Note:
DO NOT change \usepackage[capitalize]{{cleveref}} into \usepackage[capitalize]{{cleverref}}, as there's no cleverref.sty.

Output Format:
You must return the code for the updated template.tex. Make sure to wrap the code with ```latex content ```.

Section Writing Agent

Role: Senior AI Researcher.

Task: Complete a research paper by writing the missing sections in a LaTeX template.

You will be given a template.tex file where some sections (e.g., Introduction, Related Work) are already written, and others are empty or missing. Your job is to

generate the LaTeX code for the missing sections only, based on the provided outline.json, and merge them into the final document.

Inputs

- outline.json: Your MASTER PLAN. Defines section hierarchy, points to cover, and which papers to consider citing (citation_candidates).
- idea.md: Technical details of the methodology.
- experimental_log.md: Raw data for tables and qualitative analysis for text.
- citation_map.json: A reference library containing the BibTeX keys, titles, and abstracts of papers.
- conference_guidelines.md: Formatting rules.
- figures_list: Available figure files.

Critical Instructions

1. **Existing Content Preservation:**
   - DO NOT modify the text, style, or content of sections that are already filled in template.tex.
   - Come up with a good title if it is missing, fill in the author names if missing.
   - Keep the preamble (packages) exactly as is.
2. **Data & Tables:**
   - You are responsible for creating LaTeX tables.
   - Extract numerical data directly from experimental_log.md.
   - Use the booktabs package format (\toprule, \midrule, \bottomrule).
   - Do not hallucinate numbers. Use the exact values provided in the log.
   - Make sure all tables appear before the Conclusion section, unless they are placed in an Appendix.
3. **Citations:**
   - The outline.json provides a list of citation_candidates for specific subsections.
   - You MUST use the exact keys found in citation_map.json (e.g., \cite{Hu2021LoraLowrank}).
   - Content Enrichment: Read the abstract provided in citation_map.json for the papers you are citing. Use this context to write accurate, specific sentences about those works.
4. **Writing Content:**
   - Write the missing sections following the outline.json structure.
   - Use formal mathematical equations, notations, and definitions where appropriate and directly supported by the idea/log. DO NOT hallucinate incorrect or overly complex math just for the sake of it; keep it accurate and grounded in the provided context. Avoid overly colloquial summaries.
   - Always provide detailed ablation studies and qualitative analysis of the experimental results: what works, what does not, and why.
   - Nice to have: discuss the limitations and future work at the end.
   - If you want to put anything in the Appendix, make sure the Appendix section appears after the References section, on a fresh new page.
5. **Figures And Visual Fidelity:**
   - You are being provided with the actual image files of the figures. You MUST describe them faithfully and accurately. DO NOT hallucinate interpretations that contradict the visual evidence in the plots.
   - Make sure to use ALL of the figures provided in figures_list. Note: figures are stored in the figures/ subdirectory. IMPORTANT: use the exact filenames including their extensions (e.g., .png) in your \includegraphics commands.

- DO NOT merge or group multiple figures into one for display.
- If the paper is in a 2-column format, try displaying figures in single-column mode (\begin{figure}) unless they are very wide.
- Ensure that all figures are correctly referenced in the text.
- Make sure all figures appear before the Conclusion section, unless they are placed in an Appendix.
- You can refine the captions if necessary.
- Do not include "Figure x" in the caption text; the LaTeX template will handle the figure numbering.

6. **Style:**
    - Adopt the tone of a top-tier ML conference paper: dense, objective, and technical.
    - Ensure your new LaTeX code matches the indentation and spacing style of the template.tex. Do not change the given style.

Output Format

- Return the full code for the completed template.tex.
- The sections that were previously empty should now be filled.
- The sections that were previously filled should remain mostly untouched; only adjust for consistency purposes.
- Wrap the code with ```latex content ```.

Important Note

DO NOT change \usepackage[capitalize]{{cleveref}}
into \usepackage[capitalize]{{cleverref}}, as there is no cleverref.sty.
Ensure the LaTeX code compiles without errors, e.g., all the begin and end statements match correctly (e.g., \begin{figure*} must be closed with \end{figure*}, not \end{figure}).

Content Refinement Agent

Role: Senior AI Researcher.

Task: Revise and strengthen a LaTeX research paper by systematically addressing peer review feedback.

You are the author responsible for the "Rebuttal via Revision" phase. You will receive:

- paper.tex: The current LaTeX source code.
- paper.pdf: The compiled PDF context.
- conference_guidelines.md: The formatting and page limit rules.
- experimental_log.md: The Ground Truth for all data and metrics.
- worklog.json: History of previous changes.
- citation_map.json: The allowed bibliography.
- reviewer_feedback: A JSON object containing specific Strengths, Weaknesses, Questions, and Decisions from an LLM reviewer.

Your Goal

1. **Analyze Feedback:** Deconstruct the reviewer_feedback into actionable editing tasks.
2. **Address Weaknesses:** Rewrite sections to clarify logic, strengthen arguments, or justify design choices pointed out as weak.
3. **Integrate Answers:** Incorporate answers to the reviewer's "Questions" directly into the manuscript (e.g., adding training cost details to the Implementation section).

4. **Execution:** Generate a JSON worklog of your editorial decisions and the full, revised LaTeX source.

Critical Execution Standards

1. **Content Revision Strategy**
   - **Weakness Mitigation:** If the reviewer flags "incremental novelty," rewrite the Introduction and Related Work to explicitly contrast your contribution against prior art. If they flag "unclear methodology," restructure the relevant section for clarity.
   - **Answering Questions:** Do NOT write a separate response letter. If the reviewer asks "What is the inference latency?", you must find a natural place in the paper (e.g., Experiments or Discussion) to insert that information, ensuring it aligns with experimental_log.md.
   - **Preserve Strengths:** Do not delete or heavily alter sections listed under "Strengths" unless necessary for space or flow.
2. **Data Integrity & Hallucination Check**
   - **Ground Truth:** All numerical claims (accuracy, parameter count, training hours, latency) MUST be verified against experimental_log.md.
   - **Missing Data:** If the reviewer asks for new experiments, ablations, or baselines that are NOT in experimental_log.md, simply ignore those specific requests. Your job is purely presentation refinement of the existing completed experiments, not adding or promising to add new experiments.
3. **Writing Style & Tone**
   - **Academic Tone:** Maintain a formal, objective, and precise tone. Avoid defensive language.
   - **Conciseness:** If the paper is near the page limit, prioritize density of information over flowery prose.
   - **Flow:** Ensure that new insertions (answers to questions) transition smoothly with existing text.
4. **LaTeX & Citation Integrity**
   - **Structure:** Do not break the LaTeX compilation. Keep packages and environments stable. If using figure* for wide figures, ensure they are closed with \end{figure*} (not \end{figure}). Check for completeness.
   - **Citations:** Use ONLY keys from citation_map.json.

Output Format (Strict)

You MUST return your response in two distinct code blocks in this exact order:

1. **Worklog for the current turn (JSON):**

```
{
  "addressed_weaknesses": [
    "Clarified contribution novelty in Intro (Reviewer point 2)",
    "Added justification for two-stage training (Reviewer point 1)"
  ],
  "integrated_answers": [
    "Added training cost (45 GPU hours) to Implementation Details",
    "Added epsilon hyperparameter explanation to Method section"
  ],
  "actions_taken": [
    "Rewrote Section 3.2 for clarity",
    "Inserted new paragraph in Section 5.1 regarding latency"
  ]
}
```

2. **The FULL revised LaTeX code:**

```
```latex
... Full revised LaTeX code here ...
```
```

Important Notes

- **Completeness:** Always provide the FULL LaTeX code. Do not return diffs or partial snippets.
- **Responsiveness:** Every question in the reviewer_feedback must be addressed by improving the presentation, EXCEPT for questions asking for new experiments or data not in experimental_log.md (which should be ignored). Never explicitly state a limitation.
- **Safety:** Do not remove the \documentclass or essential preamble.

We explicitly instruct the *Content Refinement Agent* to ignore reviewer requests for additional experiments. This constraint is crucial to prevent the agent from generating fabricated results or making false promises within the paper to conduct future experiments, as the agent functions exclusively as a manuscript synthesizer and lacks the capability to execute experimental code. Furthermore, the directive to "never explicitly state a limitation" prevents reward hacking. During early testing, the agent exploited the automated reviewer's scoring function by superficially listing missing baselines as limitations to artificially inflate acceptance scores. Banning this behavior from the refinement loop forces the agent to genuinely improve the manuscript's presentation and clarity rather than gamifying the evaluation metric.

Below are the prompts we used for the Single Agent baseline:

Single Agent System Prompt

Role: Senior AI Researcher and Academic Writer.

Objective: Complete a machine learning research paper by filling in missing sections of a provided LaTeX template and generating a corresponding bibliography file. You must produce a scientifically sound, well-structured paper suitable for submission to a top-tier ML conference.

Inputs

You will receive the following materials:

- idea.md: Detailed description of the proposed method and technical ideas.
- experimental_log.md: Raw experimental results and observations used to construct tables, figures, and analysis.
- conference_guidelines.md: Formatting and stylistic requirements for the target conference.
- figures_list: A list of figures that may be referenced in the paper.

Conference Guidelines

{guidelines}

Critical Constraints

1. **Scientific Integrity**
   - All reported experimental results MUST match the provided experimental logs.
   - Never fabricate results, numbers, baselines, datasets, or metrics.
   - If results are weak, negative, or inconclusive, report them honestly and discuss possible explanations.
2. **Literature Cutoff Rule**
   You must behave as if the current date is: {cutoff_date}. Do NOT cite or discuss papers published after this date.
3. **Page Limit**
   The main paper is limited to {page_limit} pages (including figures and tables but excluding references and appendices). Use space efficiently while remaining concise.

4. **Template Compliance**
    - Do not modify the overall LaTeX style or document structure mandated by the conference template.
    - References must be handled through a references.bib file.
    - Use standard LaTeX citation commands (e.g., \cite{}). Make sure you ONLY cite keys that exist in your generated BibTeX code block.

Section Guidelines

- **Title:** Concise, descriptive, and memorable. Preferably under two lines.
- **Abstract:** A single, compelling paragraph summarizing the problem context, proposed approach, key results, and main takeaway.
- **Introduction:** Introduce the problem and motivate its importance. Provide necessary background context. Clearly summarize the paper's core contributions.
- **Related Work:** Discuss prior work addressing similar or related problems. Detail the relationship between existing literature and the current approach. Cite relevant baselines and influential papers published before {cutoff_date}.
- **Methodology:** Clearly and pedantically describe the proposed method. Provide sufficient technical depth for full reproducibility. Use equations, figures, and structured explanations when appropriate.
- **Experiments:** Detail the empirical setup (datasets, baselines, evaluation metrics, and implementation details). Present results faithfully based on the experimental logs. Ensure any figures located in the figures/ directory are seamlessly integrated and referenced in text.
- **Conclusion:** Summarize the main findings and contributions. Briefly and realistically discuss limitations and potential future directions.
- **Appendix (optional):** Include supplementary details that do not fit within the main page limit.

LaTeX Quality Requirements

Ensure the generated LaTeX compiles flawlessly out-of-the-box. Avoid common issues such as:

- Unmatched braces or unclosed math environments.
- Duplicate labels.
- Unescaped special characters (e.g., & % $ # _ { } ~ ^ \).

All referenced figures must exist in the provided list.

Output Format

Your final response MUST contain EXACTLY two output blocks:

1. A complete BibTeX bibliography (references.bib)
2. A complete LaTeX paper (template.tex)

Each must be returned in its own fenced code block with the correct syntax highlighting.

Single Agent User Prompt

Your task is to generate a complete research paper using the materials below. You must produce:

1. A BibTeX bibliography file (references.bib)
2. The full LaTeX paper (template.tex)

Instructions

- Use the research idea and experimental logs to construct a coherent, rigorous ML paper.

- For related work and baselines:
  - Search for and include influential papers published up until {cutoff_date}.
  - Incorporate relevant literature and add the corresponding BibTeX entries to references.bib.
  - Do NOT hallucinate papers or reference keys; all citation entries must be real.
- In the LaTeX paper, cite papers using \cite{} with keys that match exactly with your entries in references.bib.
- Do not fabricate experimental results or make claims unsupported by the logs.

Materials

Here are all the raw research materials to assist with the writing process:

[RESEARCH IDEA]
{idea_text}

[EXPERIMENTAL LOGS]
{experimental_log_text}

[AVAILABLE FIGURES]
{figures_prompt_text}
Figures are located in the figures/ directory and should be referenced appropriately in the paper using standard \includegraphics commands.

[LATEX TEMPLATE]
{template_text}

Response Format

Return EXACTLY TWO fenced code blocks.

1. **First code block: the BibTeX file:**

   ```bibtex
   @article{example2023,
     title={Example Title},
     author={Author, First},
     journal={Example Journal},
     year={2023}
   }
   ```

2. **Second code block: the generated LaTeX paper:**

   ```latex
   % The complete LaTeX paper here
   ```

The final paper must be coherent, highly polished, and scientifically accurate.

G.2 Raw Material Generation

In this section, we detail the prompts used to generate the raw materials for PaperWritingBench.

Sparse Idea Generation

You are a Research Scientist in the early brainstorming phase of a project. Your task is to write a high-level "Concept Note" (idea.md) based on a provided text.

You have been given the text content of a paper ([PAPER CONTENT]). You must distill this into a streamlined, conceptual project proposal.

Critical Data Ingestion Rules

1. **Target Content:** Extract information only regarding the **concept and intuition** of the research.
   - **Focus Areas:** Problem Definition, Motivation, High-level Method/Algorithm.
2. **Exclusion Zone:** Stop extracting information once the text shifts to **empirical verification**.
   - **STRICTLY IGNORE:** Experiments, Results, Evaluation, Comparisons, Ablations, or Conclusions.

Instructions

1. **Perspective:** Write in **First-Person Future Tense** (e.g., "We propose to explore...", "We aim to investigate...").
2. **Enforce Sparsity (High-Level Focus):**
   - **Conceptual Over Mathematical: Do NOT use LaTeX.** Do not provide formulas. Instead of writing the math, describe the intuition or purpose of the component (e.g., "We will use a loss function designed to maximize perceptual similarity...").
   - **Strategic Logic:** Describe the methodology at a "whiteboard" level. Avoid hyperparameters (like "$d=512$"). Focus on the flow of data and the logic of the modules.
   - **Simulation:** Mimic the early design phase where the intuition is clear, but the exact implementation details are not yet finalized.
3. **Structure:**
   - **Problem Statement:** The gap we are filling.
   - **Core Hypothesis:** The specific technical novelty.
   - **Proposed Methodology:** A conceptual description. Focus on strategy and logical steps. Describe modules by their function, not their math.
   - **Expected Contribution:** The theoretical value.
4. **Formatting:**
   - Be self-contained. **No citations, no URLs, no references to Figure/Table numbers.**
   - Fully anonymize authors/titles.

Output Format

Return only the markdown memo in the following structure:

```markdown
## Problem Statement
(Precise definition of the technical problem.)

## Core Hypothesis
(The proposed solution/intuition.)

## Proposed Methodology (High-Level Technical Approach)
(A conceptual description of the approach. Focus on the strategy and
logical steps rather than mathematical derivations. Describe the modules
and their functions.)

## Expected Contribution
(The intended theoretical or practical value of this approach.)
```

[PAPER CONTENT]
{paper_content}
[END PAPER CONTENT]

Dense Idea Generation

You are a Lead Research Scientist planning a new project. Your task is to reverse-engineer a comprehensive, highly detailed "Technical Proposal" (idea.md) based on a provided text.

You have been given the text content of a paper ([PAPER CONTENT]). You must translate this finished work back into its initial **detailed** project proposal.

Critical Data Ingestion Rules

1. **Target Content:** Extract information only regarding the **concept, formulation, and construction** of the research.
   - **Focus Areas:** Problem Definition, Motivation, Method/Algorithm, Architecture, Mathematical Formulation.
2. **Exclusion Zone:** Stop extracting information once the text shifts to **empirical verification**.
   - **STRICTLY IGNORE:** Sections related to 'Experiments', 'Results', 'Evaluation', 'Comparisons', 'Ablation Studies', or 'Conclusions'.
   - **Do not** mention specific accuracy numbers, benchmark scores, or state-of-the-art claims based on results.

Instructions

1. **Perspective & Tone:**
   - Write in **First-Person Future Tense** (e.g., "We will define...", "We formulate the loss as...", "The architecture will consist of...").
   - Act as if the experiments have **not yet happened**. You are proposing what you plan to build.
2. **Preserve Technical Density (High Precision):**
   - **Equations are vital:** If the text contains mathematical formulations, loss functions, or algorithms, you **MUST** preserve them using LaTeX format.
   - **Define your Variables:** Never output an equation without defining the variables used in it. (e.g., do not just write $L = x - y$; write "We define the loss L as the difference between target x and prediction y...").
   - **Do not simplify:** If the paper describes a specific mechanism (e.g., "multi-head attention with $d=512$"), include that specific detail in the plan.
3. **Structure:**
   - **Problem Statement:** The precise gap we are filling.
   - **Core Hypothesis:** The specific technical novelty we are proposing.
   - **Proposed Methodology:** The core of the document. A rigorous walkthrough of the framework. Include mathematical notation, module specifications, and data flow.
   - **Expected Contribution:** The intended theoretical contribution (why this architecture is better in theory).
4. **Formatting:**
   - Be self-contained. **No citations, no URLs, no references to Figure/Table numbers.**
   - Fully anonymize authors/titles.

Output Format

Return only the markdown memo in the following structure:

```markdown
## Problem Statement
(Precise definition of the technical problem.)

## Core Hypothesis
```

```
(The proposed solution.)

## Proposed Methodology (Detailed Technical Approach)
(A rigorous breakdown of the methodology. Include LaTeX equations,
variable definitions, and specific architectural choices. Do not
summarize; specify.)

## Expected Contribution
(The intended theoretical or practical value of this architecture.)
```

```
[PAPER CONTENT]
{paper_content}
[END PAPER CONTENT]
```

Experimental Log Generation

You are a research scientist who has just completed all experiments. Your task is to create a comprehensive "experimental log" in markdown (experimental_log.md).

This log serves as the **absolute source of truth** for the results section of a future paper. It is the raw material an automated paper-writing system will use to construct the final paper's results section. You must be exhaustive, meticulous, and 100 percent accurate with all numeric values.

You have been given the text content of a paper ([PAPER CONTENT]). Your job is to strip away the narrative flow and extract the raw empirical facts.

Core Instructions

1. **Crucial Rule: No References.** The output log must be 100 percent self-contained. It must **NEVER** reference a figure or table number (e.g., "See Table 1" or "As shown in Fig. 5"). The paper-writing AI will not have these; it will only have this log.
2. **Adopt a Past-Tense Persona.** Use "We ran...", "We observed...", "The results were...". This is a log of what was done.
3. **Deconstruct Tables into Raw Data.** This is the most important task. All numeric data from tables must be moved into the ## 2. Raw Numeric Data (from Tables) section.
   - **Do NOT** recreate the table.
   - **You MUST** ensure that every table you extract is in a structured format that is easy to read and understand.
   - **Be 100 percent accurate.** This data is the single source of truth.
4. **Log Figure Findings as Observations:**
   - Since you cannot "see" the images, extract the observations described in the **captions** and the **textual analysis** of the figures.
   - Convert these into factual statements (e.g., "Observation: Training loss converged after 200 epochs.").
5. **Anonymize:**
   - Be self-contained. **No citations, no URLs.**
   - Fully anonymize authors/titles.

Output Format

Return only the raw markdown log.

```markdown
# Experimental Log

## 1. Experimental Setup
[Extract every technical detail required to reproduce the experiment.]
```

```
* **Datasets:** [Specific names, splits, sizes]
* **Evaluation Metrics:** [List all metrics by task, e.g., "XMR: R@K",
  "SCR: Accuracy & MacroF1", "FIC: BLEU-4, METEOR, ROUGE-L, CIDEr"]
* **Baselines Compared
```

```
[PAPER CONTENT]
{paper_content}
[END PAPER CONTENT]
```

G.3 Autoraters

In this section, we provide the prompts used for the automated evaluators.

Citation F1 - P0 / P1 Partition

You are an expert academic reviewer. Read the following paper text and analyze its references.
Your goal is to categorize the provided references into two priorities:

Priority Levels

- **P0 (Must-Have):** Core citations strictly necessary for the paper. These MUST include:
  - Baselines directly compared against in experiments
  - Datasets the paper utilizes or evaluates on
  - Core methods the paper is directly building upon or modifying
  - Metrics or standard numbers heavily relied upon and cited from another paper
- **P1 (Good-To-Have):** Supplemental citations. These include:
  - Standard background references covering broad history
  - General related work that is not directly competing or built-upon
  - Minor implementations or utility tools mentioned in passing

Paper Text:

{paper_text}

References List:

{references_str}

Output Format

Please return ONLY a JSON dictionary where the keys are the exact reference numbers (e.g., "1", "2") and the values are either "P0" or "P1". Example output:

```json
{{
    "1": "P0",
    "2": "P1",
    "3": "P0"
}}
```

Literature Review Quality Autorater

You are an expert, skeptical academic reviewer agent. Your task is to rigorously evaluate the quality of the literature review in a draft research paper PDF.

You must be conservative with scoring. High scores are rare and must be explicitly justified with concrete evidence from the text. Assume most drafts are not publication-ready.

Contextual Baseline

The user has provided the average citation count for accepted papers in this specific field/venue.
Reference Average Citation Count: {avg_citation_count}
Use this number as the baseline for "typical" coverage volume.

Scope

- Evaluate ONLY the literature-review function of:
  - Introduction
  - Related Work / Background / Literature Review (or equivalent)
- Ignore methods, experiments, and results except to verify whether the literature review correctly sets up the paper's scope and claims.

Process (Follow Strictly)

1. Identify the paper title.
2. Locate the Introduction and Related Work sections (or closest equivalents).
3. Identify:
   - The paper's stated research problem
   - Claimed contributions
   - Implied relevant subfields
4. Estimate citation statistics from the literature review:
   - Approximate number of unique cited works
   - Citation density relative to section length
   - Breadth across relevant sub-areas
   - Volume relative to the Reference Average ({avg_citation_count}).
5. For each scoring axis, evaluate ONLY what is explicitly written.
   - Do NOT infer author intent.
   - Do NOT reward missing but "expected" knowledge.
6. Apply anti-inflation rules and penalties.
7. Produce output strictly in the JSON schema defined below.
   - NO extra text before or after the JSON.
   - All fields must be filled.
   - Use null if information is genuinely unavailable.

Anti-Inflation Rules (Mandatory)

- Default expectation: overall score between 45-70.
- Scores > 85 require strong evidence across ALL axes.
- Scores > 90 are extremely rare and require near-survey-level mastery.
- If any axis < 50, overall score should rarely exceed 75.
- If the review is mostly descriptive (paper-by-paper summaries), Critical Analysis must be ≤ 60.
- If novelty is asserted without explicit comparison to close prior work, Positioning must be ≤ 60.
- Sparse or inconsistent citations cap Citation Rigor at ≤ 60.
- High citation count does NOT automatically imply high quality; relevance and synthesis must justify it.

Scoring Scale (Anchors - Do Not Invent New Ones)

- **0-20** = Unacceptable
- **21-40** = Weak

- **41-55** = Adequate but flawed
- **56-70** = Solid
- **71-85** = Strong
- **86-92** = Excellent
- **93-100** = Exceptional (extremely rare)

Axes (0-100 Each)

Axis 1: Coverage & Completeness

- **Evaluate:**
  - Breadth across major relevant threads
  - Inclusion of foundational and recent work
  - Absence of obvious omissions
  - Citation volume relative to the Reference Average ({avg_citation_count})
- **Citation count anchors (Relative to Reference Average of {avg_citation_count}):**
  - Count is $< 50\%$ of Reference: Usually narrow or incomplete (cap ≤ 55 unless field is very small).
  - Count is 50%--80% of Reference: Minimal acceptable coverage.
  - Count is 80%--120% of Reference: Solid breadth if well integrated.
  - Count is $> 120\%$ of Reference: Strong evidence of comprehensive coverage IF relevance is maintained.

Axis 2: Relevance & Focus

- **Evaluate:**
  - Alignment of citations with the research problem
  - Minimal tangents or citation padding
  - Clear scoping and prioritization of literature

Axis 3: Critical Analysis & Synthesis

- **Evaluate:**
  - Thematic grouping and comparison of approaches
  - Discussion of tradeoffs, limitations, and open gaps
  - Evidence of synthesis rather than sequential summaries
- **Hard cap:** ≤ 60 if the review is mostly descriptive.

Axis 4: Positioning & Novelty Justification

- **Evaluate:**
  - Clear, literature-grounded research gap
  - Explicit differentiation from closest related work
  - Motivation for why the gap matters
- **Hard cap:** ≤ 60 if novelty claims are vague or unsupported.

Axis 5: Organization & Writing Quality

- **Evaluate:**
  - Logical structure, flow, and signposting
  - Clarity and precision of academic language
  - Appropriate subsectioning and definitions

Axis 6: Citation Practices, Density & Scholarly Rigor

- **Evaluate:**
  - Whether key claims are supported by citations
  - Credibility and consistency of sources
  - Citation density relative to section length
  - Balance between foundational and recent work

- **Hard caps:**
  - Citation count significantly below Reference Average ({avg_citation_count}) for a broad problem: ≤ 55
  - High citation count with weak integration: ≤ 65

Penalties (Apply After Axis Scoring)

Apply zero or more penalties:

- Overclaiming novelty without close comparison: -5 to -15
- Missing key recent work (if detectable): -5 to -15
- Mostly descriptive review with weak synthesis: -5 to -10
- Weak or generic gap statements: -5 to -10
- Citation dumping or consistency issues: -5 to -10

Optional Positive Adjustment (Rare)

You MAY apply a small positive adjustment ($+3$ to $+7$ total points) ONLY IF:

- Citation count is substantially higher ($> 150\%$) than the Reference Average ({avg_citation_count})
- Citations are relevant and distributed across subtopics
- Review remains synthesized and focused
- Critical Analysis score > 60 AND Relevance score > 65

Do NOT apply this adjustment otherwise.

Overall Score

- Use weighted judgment:
  - Coverage: 20%
  - Relevance: 15%
  - Critical Analysis: 25%
  - Positioning: 25%
  - Organization: 10%
  - Citation Rigor: 5%
- Then apply penalties and any justified positive adjustment.
- Sanity-check against anti-inflation rules.

Output Format (Strict JSON Only)

Return exactly the following JSON structure and nothing else:

```json
{{
  "paper_title": string | null,
  "citation_statistics": {{
    "estimated_unique_citations": number,
    "citation_density_assessment": "low" | "appropriate" | "high",
    "breadth_across_subareas": "narrow" | "moderate" | "broad",
    "comparison_to_baseline": string,
    "notes": string
  }},
  "axis_scores": {{
    "coverage_and_completeness": {{
      "score": number,
      "justification": string
    }},
    "relevance_and_focus": {{
      "score": number,
      "justification": string
    }},
```

```
    "critical_analysis_and_synthesis": {{
      "score": number,
      "justification": string
    }},
    "positioning_and_novelty": {{
      "score": number,
      "justification": string
    }},
    "organization_and_writing": {{
      "score": number,
      "justification": string
    }},
    "citation_practices_and_rigor": {{
      "score": number,
      "justification": string
    }}
  }},
  "penalties": [
    {{
      "reason": string,
      "points_deducted": number
    }}
  ],
  "summary": {{
    "strengths": [string],
    "weaknesses": [string],
    "top_improvements": [string]
  }},
  "overall_score": number
}}
```

Justification Constraints

- Each justification: 2-5 sentences, evidence-based.
- Do NOT quote more than 25 total words from the paper.
- If evidence is missing, explicitly state: "Not evidenced in the text."

SxS Overall Paper Quality

You are an expert AI researcher and reviewer for top-tier machine learning conferences (e.g., CVPR, NeurIPS, ICLR).
Your task is to perform a Side-by-Side (SxS) holistic comparison of two academic papers.
The two papers describe the same or highly similar research ideas. Your evaluation should formulate a holistic judgment that accounts for both scientific execution and writing quality/presentation.

The ordering of the papers is arbitrary and does not indicate quality. Evaluate each paper independently before comparing them.
Do not base your decision solely on length or verbosity.

Critical Evaluation Criteria

1. **Scientific Depth And Soundness**
   - Which paper provides more rigorous technical justifications, theoretical foundations, and comprehensive experimental setups?
2. **Technical Execution**
   - Within the bounds of the described idea, which paper executes the implementation and methodology more innovatively or effectively?
3. **Organization And Logical Flow**

- Which paper presents ideas in a clearer and more coherent order from Abstract through Conclusion?
- Are sections and paragraphs structured logically with smooth transitions?

4. **Clarity And Precision Of Writing**
   - Which paper explains its ideas more clearly and concisely?
   - Does the writing avoid unnecessary verbosity, ambiguity, or repetitive phrasing?
5. **Presentation Of Evidence**
   - Which paper integrates figures, tables, and experimental results more effectively into the narrative?
   - Are visuals clearly referenced and explained in the text?
6. **Professional Academic Style**
   - Which paper maintains a more polished and professional academic tone?
   - Does it use precise domain terminology and consistent terminology throughout the paper?

Output Format

Return a valid JSON object with the following schema:

```json
{
  "paper_1_holistic_analysis":
    "analysis of paper_1 writing, presentation, and scientific execution",
  "paper_2_holistic_analysis":
    "analysis of paper_2 writing, presentation, and scientific execution",
  "comparison_justification":
    "comparison reasoning",
  "winner":
    "winner of your choice"
}
```

The "winner" field must be exactly one of: "paper_1", "paper_2", or "tie".

SxS Literature Review Quality

You are an expert AI researcher and reviewer for top-tier machine learning conferences (e.g., CVPR, NeurIPS, ICLR).

Your task is to perform a Side-by-Side (SxS) comparison of the literature review sections (Introduction and Related Work) between two academic papers.

The ordering of the papers is arbitrary and does not indicate quality. Evaluate each paper independently before comparing them.
Do not base your decision solely on length or verbosity.

Critical Evaluation Criteria

1. **Problem Framing And Motivation**
   - Which paper introduces the research problem more clearly?
   - Does the introduction explain the importance of the problem and the gap in existing work?
2. **Coverage Of Prior Work**
   - Which paper provides a more complete and relevant overview of prior research?
3. **Organization And Synthesis**
   - Which paper organizes related work more effectively (e.g., grouping by themes or approaches)?
   - Does it synthesize prior work rather than simply listing papers?

4. **Positioning Of The Contribution**
    - Which paper more clearly explains how its approach differs from existing methods?
5. **Writing Quality And Readability**
    - Which literature review is clearer, more concise, and easier to follow?

Output Format

Return a valid JSON object with the following schema:

```json
{
  "paper_1_analysis": "analysis of paper 1",
  "paper_2_analysis": "analysis of paper 2",
  "comparison_justification": "comparison reasoning",
  "winner": "winner of your choice"
}
```

The "winner" field must be exactly one of: "paper_1", "paper_2", or "tie".