TGIF: Text-Guided Layer Fusion Mitigates Hallucination in Multimodal LLMs

Chenchen Lin[1], Sanbao Su[1,*], Rachel Luo[2], Yuxiao Chen[2], Yan Wang[2], Marco Pavone[2,3,†], Fei Miao[1‡]

[1]University of Connecticut [2]NVIDIA [3]Stanford University

{chenchen.lin, sanbao.su, fei.miao}@uconn.edu
{raluo, yuxiaoc, yanwan, mpavone}@nvidia.com

Abstract

Multimodal large language models (MLLMs) often generate confident but visually ungrounded responses, a phenomenon known as hallucination. While prior mitigation methods primarily operate at decoding time or rely on additional training, they typically expose the language model to a fixed, late-layer visual representation, overlooking the rich hierarchy encoded by vision transformers. We show that hallucination behavior is strongly influenced by the depth of visual features provided to the LLM, and that no single layer is optimal across queries. To address this, we propose Text-Guided Inter-layer Fusion (TGIF), a lightweight architectural module that dynamically reweights visual features across transformer layers based on the input text, without modifying the vision encoder or increasing the token budget. Experiments on hallucination, OCR, and general VQA benchmarks demonstrate that TGIF substantially improves visual grounding and hallucination robustness while preserving strong overall reasoning performance. **Code:** github.com/Linchenchen/TGIF.

1 Introduction

Multimodal large language models (MLLMs) combine the reasoning capabilities of large language models (LLMs) with pretrained vision encoders, achieving impressive results in diverse visual tasks Liu et al. (2023c); Li et al. (2023b); Dai et al. (2023b). However, MLLMs face persistent hallucination challenge, generating confident but ungrounded content that is plausible under language priors yet inconsistent with the image Liu et al. (2024b). This is particularly pronounced in detail-oriented tasks where high-level semantic features alone are insufficient for precise grounding. Most existing hallucination mitigation methods intervene after training or at inference time Gunjal et al. (2024); Yu et al. (2024); Leng et al. (2023); Liu et al. (2024c); Huang et al. (2024); Tang et al. (2025); Chen et al. (2025a). While effective, these approaches primarily operate on the text generation process and leave the underlying visual representations unchanged.

A key but underexplored factor is the depth of visual representations exposed to the language model. Vision transformers such as CLIP encode a hierarchy of visual abstractions across layers, ranging from low-level spatial cues to high-level semantic concepts. However, most MLLMs expose only a single, fixed (typically late) vision layer to the LLM via an MLP projector. Figure 1 presents a representative example in which the baseline LLaVA model hallucinates, demonstrating that injecting visual features from different depths, with all other components held fixed, results in qualitatively different behaviors ranging from under-recognition to hallucination.

*Current affiliation: Google.
†Corresponding author.
‡Current affiliation: Zhejiang University. Corresponding author.

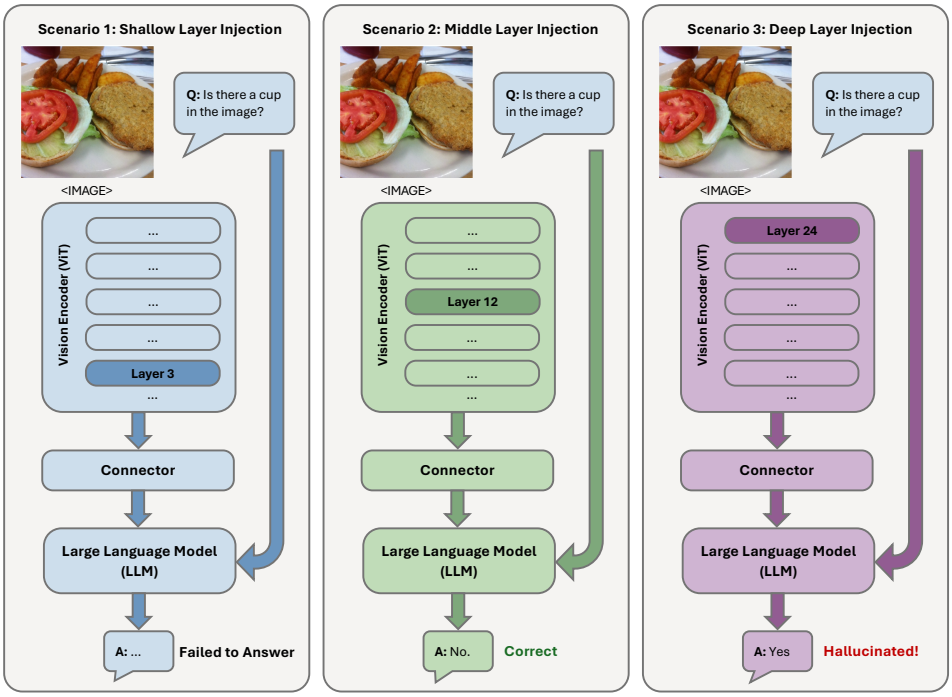

Figure 1: **Impact of ViT Layer Injection on Hallucination Behavior.**

Recent works begin to highlight the role of visual encoder layers in MLLMs, suggesting that shallow and middle layers exhibit unique strengths over deep layers in certain subtasks Chen et al. (2025b). VaLiD Liu et al. (2024c) and LayerCD Chen et al. (2025a) collectively suggest that hallucination behavior is sensitive to the choice of visual feature depth. To the best of our knowledge, however, no prior work has addressed this issue by explicitly modifying the multimodal architecture to dynamically control which visual representations are exposed to the LLM.

Motivated by this observation, we hypothesize that hallucination in MLLMs is strongly influenced by how visual features are selected and exposed to the language model, and that adapting this selection based on the input prompt can improve visual grounding. We propose **TGIF** (**T**ext-**G**uided **I**nter-layer **F**usion), a lightweight architectural module that dynamically reweights visual features across all layers of a frozen vision encoder according to the textual query (Fig. 2). This design preserves the pretrained backbone and token budget while enabling prompt-dependent control over visual abstraction.

We instantiate TGIF on top of LLaVA-1.5 and evaluate it on hallucination, OCR, and general VQA benchmarks. TGIF consistently improves hallucination robustness and fine-grained visual perception while maintaining competitive performance on general reasoning tasks, demonstrating that better visual grounding can be achieved through dynamic representation-level control rather than post hoc decoding intervention. Our main contributions are three-fold:

- We identify a limitation of current MLLMs: visual tokens are typically drawn from a single late-layer representation, which is poorly suited for detail-sensitive grounding and exacerbates hallucination under strong language priors.
- We propose **TGIF**, a text-guided inter-layer fusion module that dynamically reweights CLIP layers per query while remaining parameter- and token-efficient.
- We show that TGIF improves hallucination robustness and fine-grained visual perception on POPE, HallusionBench, OCRBench, and TextVQA, while maintaining competitive performance on general reasoning benchmarks.

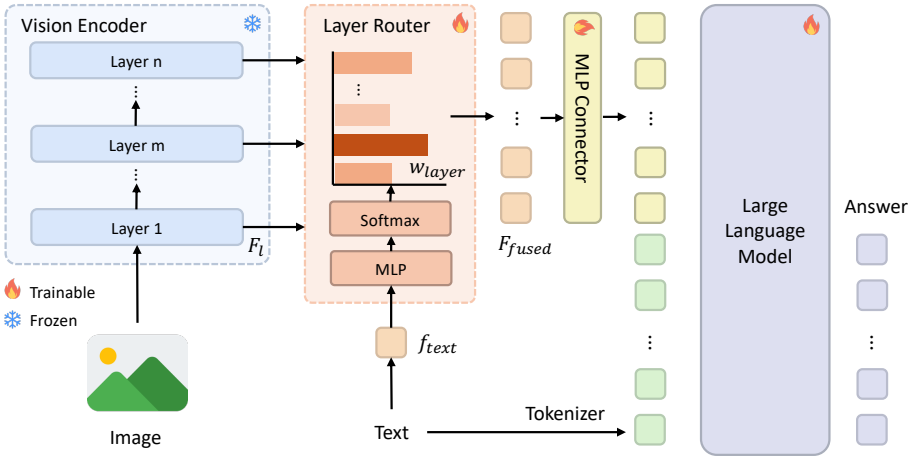

Figure 2: **Overview of the proposed Text-Guided Inter-Layer Fusion (TGIF) framework.** TGIF dynamically integrates multi-layer visual features $\{F_l\}$ from a frozen ViT based on the textual query. A Layer Router processes the text embedding f_{text} to produce a soft distribution w_{layer} via an MLP and softmax. These weights fuse hierarchical representations into F_{fused}, which is then projected and concatenated with text tokens for LLM reasoning.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) typically combine a frozen vision encoder with a language model through lightweight connectors Liu et al. (2023c); Li et al. (2023b); Dai et al. (2023b). Early designs adopt simple MLP projectors Liu et al. (2023c), while later approaches introduce query-based extraction mechanisms Dai et al. (2023b). Despite these architectural differences, most MLLMs rely on a *fixed* visual interface, commonly projecting features from a single late layer of the vision encoder. This design implicitly assumes that a single level of visual abstraction is sufficient across all queries, leaving limited flexibility to adapt to diverse grounding requirements.

2.2 Hallucination and Representation Depth

Hallucination in MLLMs, where models generate visually ungrounded content, has been widely studied Rohrbach et al. (2019). Existing mitigation strategies largely operate at the decoding or training level, including RLHF-based alignment Gunjal et al. (2024); Yu et al. (2024) and decoding-time interventions such as VCD, OPERA, FarSight and SHIELD Leng et al. (2023); Huang et al. (2024); Tang et al. (2025); Huang et al. (2026). Both LayerCD Tong et al. (2025) and VaLiD Wang et al. (2025) investigate the role of vision encoder layers by coupling hierarchical feature exploration with contrastive decoding strategies. However, recent work suggests that such improvements may not always correspond to genuine grounding gains Yin et al. (2025), motivating investigation into the representation stage.

2.3 Multi-Layer Feature Integration

Transformer-based vision encoders exhibit a hierarchical structure, where shallow and intermediate layers preserve spatial details while deeper layers capture semantic abstractions Radford et al. (2021); Gandelsman et al. (2024). Prior work has explored integrating multiple layers to enrich visual representations. For example, DenseConnector Yao et al. (2024) concatenates features from different depths, while MMFuser Cao et al. (2024) retrieves fine-grained information using deep-layer queries. More recent analyses further suggest that exposing multiple layers alone is insufficient, and that the effectiveness of different layers depends on the task and query Lin et al. (2025). Building on these insights, TGIF

performs text-guided inter-layer fusion, dynamically reweighting visual features according to the input query to improve grounding and reduce hallucination.

3 Method

Our approach is situated within a standard Vision Language Model (VLM) architecture, which comprises a vision encoder (e.g., CLIP ViT Radford et al. (2021)), a language model (e.g., Vicuna Chiang et al. (2023)), and a multimodal projector that maps visual features into the language model's embedding space. Our proposed text-guided layer selection module is designed as a core component of this multimodal projector. It takes the full stack of hidden states from the all vision encoder layers as input, along with features derived from the text prompt, to produce a dynamically tailored visual representation for the LLM.

3.1 Text-Guided Layer Selection

As established in prior work Chen et al. (2025b), different CLIP layers capture distinct semantic information, with shallow layers encoding textures and spatial details while deeper layers align more with global semantics. This suggests that the layers can be viewed as a pool of specialized "experts". Inspired by the Mixture-of-Experts (MoE) paradigm, we treat each layer as an expert and propose a dynamic layer selection framework, called TGIF (Text-Guided Inter-layer Fusion). In this framework, a text-guided "router" learns to generate weights to select and fuse the most relevant layer experts for a given task. We explore two architectures for the TGIF router, one with text-only input and other with multimodal input. The model architecture and framework is shown in Fig 2. This design allows adaptive, query-conditioned fusion of depth-wise visual cues, improving both grounding and fine-grained detail understanding.

3.1.1 Text-Guided MLP Router

Our baseline approach uses a lightweight MLP-based router to predict layer importance scores based solely on the textual prompt. This router learns a direct mapping from the question's semantics to the relevance of different vision layers. Let $\mathbf{f}_{\text{text}} \in \mathbb{R}^{D_t}$ be the pooled text embedding from the LLM, $\{\mathbf{F}_l \in \mathbb{R}^{P \times D_v}\}_{l=1}^{L}$ be the set of patch-level visual features from all L layers of the vision encoder. The MLP selector first predicts unnormalized logits for each layer:

$$\mathbf{z} = \text{MLP}(\mathbf{f}_{\text{text}}) \in \mathbb{R}^{L}. \tag{1}$$

These logits are then transformed into a probability distribution using the softmax function to represent the layer weights:

$$\boldsymbol{w} = \text{softmax}(\mathbf{z}) \in \mathbb{R}^{L}. \tag{2}$$

The final fused visual representation is computed as a weighted sum of all layer features, where the weights are broadcast across the patch and feature dimensions:

$$\mathbf{F}_{\text{fused}} = \sum_{l=1}^{L} w_l \cdot \mathbf{F}_l \in \mathbb{R}^{P \times D_v}. \tag{3}$$

3.1.2 Multimodal MLP Router

To address cases in pretraining where the text prompt is generic (e.g., "Describe the image"), we enhance the router with visual context. This multimodal approach allows the selection to be conditioned on both the question and the image content. We first extract a global image representation, $\mathbf{f}_{\text{image}} \in \mathbb{R}^{D_v}$, by taking the [CLS] token from the penultimate layer of the vision encoder. The text and image features are then projected to a common dimension D_p and concatenated:

$$\mathbf{f}_{\text{multi}} = [\mathbf{f}_{\text{text}}\mathbf{W}_t, \mathbf{f}_{\text{image}}\mathbf{W}_v] \in \mathbb{R}^{2D_p}, \tag{4}$$

where $\mathbf{W}_t \in \mathbb{R}^{D_t \times D_p}$ and $\mathbf{W}_v \in \mathbb{R}^{D_v \times D_p}$ are learnable projection matrices. This combined multimodal feature vector is then used by the MLP to predict the layer weights, following the same process as in Equations 1-3. A schematic comparison between the text-only and multimodal routers is provided in Appendix C (Fig. 7).

3.2 Load Balancing Loss

A common challenge when training MoE-style routers is the tendency for the router to converge to a state where it consistently selects the same few "safe" experts (in our case, layers), leading to "expert starvation" Shazeer et al. (2017). To mitigate this and encourage the router to utilize a more diverse set of layers, we incorporate a modified auxiliary load balancing loss into our total training objective. For our soft-selection routers, we use an entropy-based loss. Let $\boldsymbol{w}_b \in \mathbb{R}^L$ be the layer weights for the b-th sample in a batch of size B. We first compute the average weight for each layer across the batch:

$$\bar{\boldsymbol{w}} = \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{w}_b \in \mathbb{R}^L. \tag{5}$$

The auxiliary loss is then formulated to maximize the entropy of this average distribution, which encourages a more uniform usage of layers:

$$\mathcal{L}_{\text{aux}} = \lambda \sum_{l=1}^{L} \bar{w}_l \log(\bar{w}_l + \epsilon), \tag{6}$$

where λ is a hyperparameter controlling the strength of the loss and ϵ is a small constant for numerical stability. This auxiliary loss is added to the main VLM loss during training. Because the nature of textual prompts differs between pretraining and instruction tuning dataset, we apply different λ values across stages. During pretraining, the router often receives generic prompts (e.g., "Describe the image"), which provide limited textual guidance. We thus strengthen the visual signal by applying a slightly larger λ to encourage exploration of multiple layers. During fine-tuning, prompts are task-oriented and semantically rich (e.g., "What number is written on the sign?"). So we reduce λ to allow the router to focus on discriminative, text-conditioned layer selection.

4 Experiments

4.1 Experimental Setting

We implement our proposed framework, TGIF, on top of the publicly available LLaVA-1.5 Li et al. (2023a) codebase. To ensure a fair comparison, our primary experiments maintain consistency with LLaVA-1.5 by employing CLIP-ViT-L/14-336px Radford et al. (2021) as the vision encoder and Vicuna-7B Chiang et al. (2023) as the LLM. The training details can be found in Sec. D.

4.2 Baseline Methods

We evaluate TGIF against two complementary categories of baselines:

Decoding-based Hallucination Mitigation. VCD Leng et al. (2023), VTI Liu et al. (2024c), OPERA Huang et al. (2024), FarSight Tang et al. (2025), PerturboLLaVA Chen et al. (2025a), SHIELD Huang et al. (2026), and RepEdit-GA Shi et al. (2025) intervene at decoding or inference time. TGIF is complementary to these methods: it is a visual feature-level architectural modification trained within the original MLLM pretraining/instruction-tuning pipeline, without requiring an additional hallucination-specific dataset, a separate hallucination-mitigation finetuning stage, or an extra inference-time decoding branch.

Multi-layer Fusion Baselines. DenseConnector Yao et al. (2024) and MMFuser Cao et al. (2024) aggregate features from multiple vision encoder layers using static fusion strategies.

4.3 Quantitative Results

Hallucination Benchmarks. To assess grounding faithfulness, we adopt two representative hallucination benchmarks: HallusionBench (HB) Liu et al. (2024a) and POPE Yifan Li & Wen (2023). HallusionBench probes visual factuality by testing a model's ability to reject implausible object claims, while POPE reformulates hallucination detection as a binary classification task, quantifying the model's awareness of object existence.

OCR Benchmarks. To evaluate text recognition and detail-sensitive reasoning, we include TextVQA Singh et al. (2019b) and the comprehensive OCRBench Liu et al. (2024e), which covers scene-text, document-text, and key information extraction subtasks. These benchmarks reveal the model's ability to retrieve and interpret fine-grained textual cues—an area where hallucination often manifests due to overreliance on semantic priors.

Relation Grounding. To evaluate fine-grained relational grounding beyond object-existence recognition, we additionally evaluate TGIF on MMRel Nie et al. (2025), covering action, spatial, and comparative relations. TGIF improves overall accuracy from 61.0 to 63.0 on 7B and from 64.5 to 75.4 on 13B, with particularly strong gains on spatial-relation subsets (Table 4). Performance on comparative relations is mixed; the full per-subset breakdown is provided in the Appendix.

General Reasoning and Overall Benchmarks. For overall multimodal reasoning performance, we report results on widely used visual QA and instruction-following datasets, including MMBench (MMB) Liu et al. (2024d), ScienceQA Lu et al. (2022) and GQA Ainslie et al. (2023). Together, these benchmarks measure both the factual grounding and generalization capability of TGIF across visual domains and task types.

Method	Adversarial		Popular		Random		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Baseline								
LLaVA-1.5-7B	78.96	77.57	81.88	80.06	82.29	81.33	81.04	79.65
Decoding-Time								
+ VCD Leng et al. (2024)	80.88	81.33	85.38	85.06	87.73	87.16	84.66	84.51
+ OPERA Huang et al. (2024)	82.90	83.10	84.30	85.20	85.40	87.90	84.20	85.40
+ VTI Liu et al. (2024c)	84.80	84.10	86.90	86.00	87.80	87.60	86.50	85.90
+ FarSight Tang et al. (2025)	84.40	78.50	86.20	80.10	87.70	82.60	86.10	80.40
+ LayerCD Tong et al. (2025)	82.09	80.59	84.25	82.53	85.77	83.93	84.04	82.35
+ VaLiD Wang et al. (2025)	83.20	83.29	87.17	86.65	89.03	88.36	86.47	86.10
+ RepEdit-GA Shi et al. (2025)	81.76	82.84	86.00	86.14	89.35	89.43	85.70	86.14
+ SHIELD Huang et al. (2026)	82.50	83.60	87.40	87.60	**91.30**	**91.10**	87.00	87.40
Architectural (Ours)								
+ TGIF (Final)	**86.00**	**85.60**	**88.23**	**87.60**	89.51	89.10	**87.91**	**87.43**

Table 1: **Evaluation on POPE Benchmark.** We report Accuracy (Acc.) and F1-score across three sampling settings: *Adversarial*, *Popular*, and *Random*. TGIF consistently outperforms both the LLaVA-1.5 baseline and specialized decoding-time mitigation methods.

4.4 Router Layer Selection Dynamics

To understand how TGIF adapts visual fusion to different query types, we visualize the learned layer-weight distributions in Fig. 3. The router exhibits clear, semantically-driven routing patterns. General queries (e.g., "Describe the image.") activate a broad mixture of mid- and high-level layers, reflecting the need for holistic scene understanding. Hallucination-sensitive queries place greater weight on early layers that preserve spatial and boundary cues, which helps verify object presence rather than relying on language priors. In contrast, OCR and detail-oriented questions concentrate weight on mid-to-late layers containing rich text strokes and structural detail. These behaviors confirm that TGIF does not rely on a fixed mixture but instead performs question-aware selection of visual experts, addressing the limitations of static multi-layer fusion.

Method	Params	All Acc ↑
LLaVA-1.5	13.0B	46.94
InstructBLIP Dai et al. (2023b)	8.2B	45.26
MiniGPT5 Zheng et al. (2023)	8.2B	40.30
LLaVA-1.5	7.0B	46.90
+ VCD Leng et al. (2024)	7.0B	46.90
+ OPERAHuang et al. (2024)	7.0B	47.10
+ PerturboLLaVAChen et al. (2025a)	7.0B	47.60
+ TGIF (ours)	7.0B	**49.94**

Table 2: **HallusionBench comparison.** We compare LLaVA-1.5-7B+TGIF with similarly sized open-source models (7–13B). We report GPT4-assisted All Accuracy. TGIF attains an All Accuracy of 49.94%, outperforming LLaVA-1.5 by +3.0% and exceeding larger 13B-parameter models. Detailed results are provided in the Appendix.

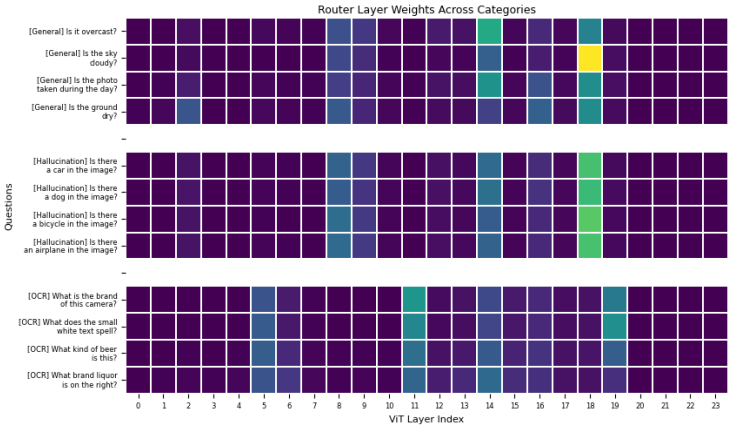

Figure 3: **Router layer selection patterns across different question categories.** This heatmap visualizes the router's learned weights for selecting vision transformer (ViT) layers across three categories of questions: **General**, **Hallucination Detection**, and **OCR/Detail Recognition**. Each row corresponds to one question, and each column indicates a specific ViT layer. Brighter colors denote higher selection weight for that layer.

4.5 Ablation Studies

4.5.1 Text-only vs. Multimodal Routing

We compare two routing strategies: text-only router that conditions solely on the question embedding, and a multimodal router that additionally incorporates a global image representation. As shown in Table 4.5, both variants consistently outperform the LLaVA-1.5 baseline, confirming the effectiveness of dynamic inter-layer fusion.

The multimodal router achieves a substantial gain on HallusionBench, suggesting that global visual context can help the router better distinguish queries requiring semantic

Method	Recog.	VQA^S	VQA^D	KIE	HMER	Final
LLaVA-1.5-7B	160	117	15	5	0	297
+TGIF (ours)	**162**	**121**	**24**	**6**	0	**313**

Table 3: **OCRBench comparison.** Recog.:Text Recognition; VQA^S:Scene Text-centric VQA; VQA^D:Document-oriented VQA; KIE: Key Information Extraction; HMER:Handwritten Math Expression Recognition.

Subset	LLaVA-1.5-7B			LLaVA-1.5-13B		
	Base	TGIF	Δ	Base	TGIF	Δ
dall-e_action	67.5	68.0	+0.5	72.8	71.3	-1.5
dall-e_spatial	51.4	52.9	+1.5	52.3	56.3	+4.0
spec_comparative	73.2	74.3	+1.1	77.4	78.5	+1.1
spec_spatial	52.2	56.6	+4.4	54.5	**68.8**	**+14.3**
vg_action	71.7	73.9	+2.2	73.6	**82.8**	**+9.2**
vg_comparative	74.2	73.7	-0.5	84.5	78.4	-6.1
vg_spatial	51.2	53.1	+1.9	56.0	**75.7**	**+19.7**
Overall	61.0	**63.0**	**+2.0**	64.5	**75.4**	**+10.9**

Table 4: **MMRel Discriminative Y/N results.** We report accuracy (%) across action, spatial, and comparative relation subsets for LLaVA-1.5-7B and LLaVA-1.5-13B. TGIF improves overall accuracy at both model scales, with particularly strong gains on spatial-relation subsets for the 13B model.

Method	Hallucination	OCR	General Reasoning		
	POPE (Acc. ↑)	TextVQA ↑	SQA ↑	GQA ↑	MMB ↑
LLaVA-1.5-7B	81.04	58.20	66.80	62.00	64.30
+ DenseConnector	86.60	**59.20**	69.50	**63.80**	66.80
+ MMFuser	86.30	58.80	68.70	62.80	**67.50**
+ TGIF (Ours)	**87.91**	58.98	**70.10**	62.58	66.40

Table 5: **Main Results on Hallucination and General Reasoning. Hallucination:** POPE; **OCR:** TextVQA; **Overall:** ScienceQA (SQA), GQA, MMBench (MMB). Best in **bold**, second best underlined.

reasoning from those demanding conservative verification. It also improves performance on ScienceQA, indicating stronger generalization on complex, multi-step reasoning tasks.

In contrast, the text-only router performs slightly better on POPE and OCRBench, where hallucination often arises from over-reliance on semantic co-occurrence rather than insufficient visual context. In these cases, conditioning routing solely on query encourages more conservative layer selection, reducing sensitivity to visually plausible but absent objects. This behavior aligns with our hypothesis that hallucination is strongly influenced by how visual abstraction depth is exposed to the language model.

Overall, while multimodal routing can benefit tasks that require global semantic understanding, we adopt the text-only router as our final configuration due to its stronger and more stable performance on hallucination-sensitive and fine-grained perception benchmarks. This choice also better isolates the effect of query-driven control over visual representation depth, which is the central focus of TGIF.

4.5.2 Visual Depth Analysis

To directly test the core hypothesis that different visual depths exhibit different grounding behavior, we construct static variants that expose only early, middle, or late ViT features, as well as a uniform fusion across layers, and evaluate them on POPE-Adversarial. As shown in Table 7, early features are conservative, achieving high precision but low recall, while late features achieve high recall at the cost of precision. No fixed depth is uniformly optimal. TGIF adaptively balances these complementary behaviors and achieves the best F1 among the depth-selection variants.

4.5.3 Effect of Load-Balancing Coefficient λ

We further analyze the impact of the entropy-based load-balancing loss on routing stability and downstream performance. As shown in Fig. 4, applying a small amount of

Router	Setting	Hallucination		OCR		Reasoning		
		POPE ↑	HB ↑	TVQA ↑	OCRB ↑	SQA ↑	GQA ↑	MMB ↑
Text-only	Vanilla (λ=0)	87.30	49.95	58.93	**31.50**	68.07	62.48	65.97
Multi.	Vanilla (λ=0)	86.26	**57.31**	**59.09**	29.90	69.72	62.37	65.46
Text-only	λ=0.01 **(Ours)**†	**87.91**	49.94	58.98	31.30	**70.10**	**62.58**	**66.40**

Table 6: **Ablation Study of TGIF Router Designs.** We compare router architectures and entropy-based load-balancing (λ). Text-only router with λ=0.01 achieves the best balance across grounding and reasoning. HB: HallusionBench; TVQA: TextVQA; OCRB: OCRBench; SQA: ScienceQA; GQA: GQA; MMB: MMBench.

Method	Acc ↑	Prec ↑	Recall ↑	F1 ↑
LLaVA-1.5-7B	0.790	0.814	0.741	0.776
Early-only	0.804	**0.963**	0.632	0.763
Mid-only	0.840	0.942	0.725	0.819
Late-only	0.837	0.800	**0.899**	0.847
Uniform	0.838	0.957	0.707	0.813
TGIF	**0.860**	0.881	0.832	**0.856**

Table 7: **POPE-Adversarial depth analysis.** Static depth choices exhibit complementary precision–recall trade-offs, while TGIF achieves the best F1 via text-guided inter-layer fusion.

regularization during pretraining only provides the best balance between VQA accuracy and hallucination robustness. In particular, the λ=0.01 pretrain-only configuration yields both the highest POPE accuracy and the strongest average VQA score, indicating that mild early-stage entropy encourages the router to explore a diverse set of layers without suppressing its ability to specialize.

In contrast, larger coefficients (e.g., λ=0.1) or applying the loss throughout full fine-tuning tend to over-regularize the router, pulling the layer distribution toward uniformity and reducing its ability to adapt the fusion pattern to the input query. These settings achieve weaker VQA performance and, in some cases, degraded hallucination resistance. Overall, the results suggest that light, pretraining-only regularization is crucial: it stabilizes routing and prevents expert collapse while still allowing the model to learn query-dependent, discriminative layer selection during instruction tuning.

4.6 Qualitative comparison

Figure 5 illustrates how text-guided inter-layer fusion alters the visual evidence exposed to the language model. The adversarial query asks whether a traffic light is present. Although the image contains no traffic light, the baseline LLaVA attends strongly to a centrally located fire hydrant whose color and shape are highly correlated with traffic lights in the training distribution. This over-reliance on late, semantically aligned visual features leads the model to hallucinate a positive answer.

In contrast, TGIF dynamically reweights visual representations across layers conditioned on the text query. The resulting relevance map deemphasizes the misleading object and distributes attention more broadly across the scene, suppressing language-driven semantic shortcuts. As a result, TGIF correctly answers "No." While this is a single example, it qualitatively demonstrates how text-guided layer fusion modulates the balance between visual abstraction and language priors, consistent with our quantitative improvements on hallucination benchmarks.

4.7 Computation and Memory Overhead

We analyze the parameter, latency, and memory overhead introduced by TGIF to validate its lightweight design. The minimal runtime overhead is expected for two reasons. First,

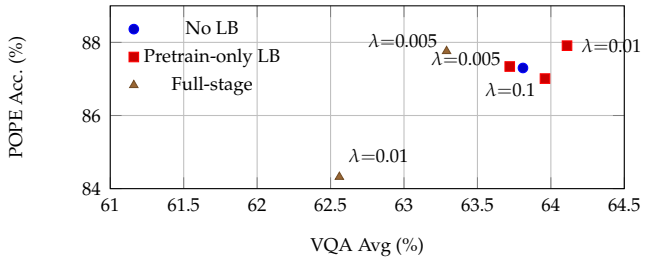

Figure 4: **Effect of load balancing on the VQA–hallucination trade-off.** Each point shows the average VQA score (ScienceQA, GQA, TextVQA) versus POPE accuracy for a different router / load-balancing configuration. We annotate all load-balancing settings next to their corresponding points.

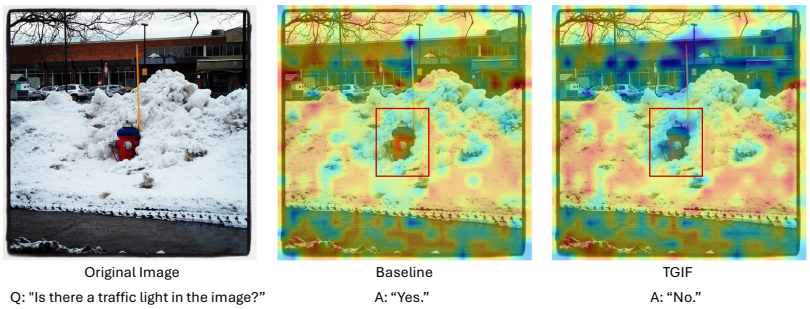

Figure 5: **Qualitative comparison on POPE adversarial. Left:** original image. **Middle:** baseline LLaVA text-conditioned relevance map. **Right:** TGIF text-conditioned relevance map. The baseline model attends strongly to a semantically misleading object (a fire hydrant resembling a traffic light), resulting in a hallucinated positive answer.

the router is a shallow MLP evaluated once per input and contributes negligible computation compared to the frozen vision encoder and the large language model. Second, LLaVA already computes and stores intermediate visual features across layers during the forward pass, so TGIF does not introduce additional vision encoder computation or memory allocation.

Model	Params (M)	Latency (s)	Peak Mem (GB)
LLaVA	7062.9	0.2203	27.16
+ TGIF	7065.0	0.2224	27.16
Overhead	+0.03%	+0.93%	+0.00

Table 8: **Computation and memory overhead.** TGIF introduces negligible overhead relative to the LLaVA-1.5 baseline.

5 Conclusion

We present TGIF, a lightweight text-guided inter-layer fusion framework that dynamically controls which visual representations are exposed to the language model. By adapting the visual abstraction level to the input query, TGIF improves visual grounding and reduces hallucination without modifying the vision encoder, increasing token count, or introducing significant computational overhead. Extensive experiments across hallucination, OCR, and general VQA benchmarks demonstrate that TGIF consistently outperforms strong baselines while preserving overall reasoning capability. These results suggest that query-adaptive control of visual feature depth is a promising direction for building more reliable and trustworthy multimodal language models.

References

Gpt-4v(ision) system card. 2023. URL https://api.semanticscholar.org/CorpusID:263218031.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023. URL https://arxiv.org/abs/2305.13245.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966, 2023. URL https://api.semanticscholar.org/CorpusID:263875678.

Yue Cao, Yangzhou Liu, Zhe Chen, Guangchen Shi, Wenhai Wang, Danhuai Zhao, and Tong Lu. Mmfuser: Multimodal multi-layer feature fuser for fine-grained vision-language understanding, 2024. URL https://arxiv.org/abs/2410.11829.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. Perturbollava: Reducing multimodal hallucinations with perturbative visual training, 2025a. URL https://arxiv.org/abs/2503.06486.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower, 2025b. URL https://arxiv.org/abs/2504.21447.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality, 2023. URL https://vicuna.lmsys.org.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023a.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*, 2023b.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition, 2024. URL https://arxiv.org/abs/2310.05916.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models, 2024. URL https://arxiv.org/abs/2308.06394.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Yiyang Huang, Liang Shi, Yitian Zhang, Yi Xu, and Yun Fu. SHIELD: Suppressing hallucinations in LVLM encoders via bias and vulnerability defense. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yk7FFLoNcP.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding, 2023. URL https://arxiv.org/abs/2311.16922.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv:2306.00890*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*, 2023b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023c. URL https://api.semanticscholar.org/CorpusID:256390509.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multi-layer visual feature fusion in multimodal llms: Methods, analysis, and best practices, 2025. URL https://arxiv.org/abs/2503.06063.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. In *CVPR*, 2024a.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models, 2024b. URL https://arxiv.org/abs/2402.00253.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023c.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. Reducing hallucinations in vision-language models via latent space steering, 2024c. URL https://arxiv.org/abs/2410.15778.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024d.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024e. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Jiahao Nie, Gongjie Zhang, Wenbin An, Yun Xing, Yap-Peng Tan, Alex C. Kot, and Shijian Lu. Mmrel: Benchmarking relation understanding in multi-modal large language models, 2025. URL https://arxiv.org/abs/2406.09121.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019. URL https://arxiv.org/abs/1809.02156.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

Youxu Shi, Suorong Yang, and Dong Liu. Exposing hallucinations to suppress them: Vlms representation editing with generative anchors, 2025. URL https://arxiv.org/abs/2509.21997.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019a.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019b. URL https://arxiv.org/abs/1904.08920.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zelin Peng, Zhiwei Yang, Jionglong Su, Minquan Lin, Yifan Peng, Xuelian Cheng, Imran Razzak, and Zongyuan Ge. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding, 2025. URL https://arxiv.org/abs/2505.16652.

Anthropic Team. Claude 3, 2024. URL https://www.anthropic.com/news/claude-3-family.

Gemini Team. Gemini: A family of highly capable multimodal models, 2023.

Bingkui Tong, Jiaer Xia, and Kaiyang Zhou. Mitigating hallucination in multimodal llms with layer contrastive decoding, 2025. URL https://arxiv.org/abs/2509.25177.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *ArXiv*, abs/2205.14100, 2022. URL https://api.semanticscholar.org/CorpusID:249152323.

Jiaqi Wang, Yifei Gao, and Jitao Sang. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding, 2025. URL https://arxiv.org/abs/2411.15839.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms, 2024. URL https://arxiv.org/abs/2405.13800.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv:2307.02499*, 2023a.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178*, 2023b.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023c.

Kun Zhou Jinpeng Wang Wayne Xin Zhao Yifan Li, Yifan Du and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=xozJw0kZXF.

Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in mllms?, 2025. URL https://arxiv.org/abs/2504.10020.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, 2024. URL https://arxiv.org/abs/2312.00849.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tongfei Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *ArXiv*, abs/2306.17107, 2023. URL https://api.semanticscholar.org/CorpusID:259287523.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *ArXiv*, abs/2310.02239, 2023. URL https://api.semanticscholar.org/CorpusID:263608981.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023a.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023b.

A Appendix

B Architectural Comparison with Prior Layer Fusion Designs

Figure 6 provides a schematic comparison between commonly used multimodal connectors and the proposed text-guided inter-layer fusion (TGIF).

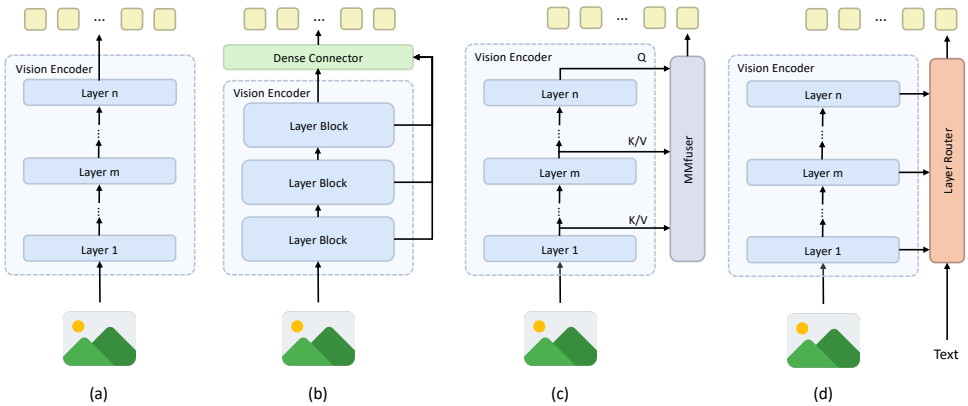

Figure 6: **Comparison of layer fusion designs in MLLMs.** (a) **MLP Connector**: Projects visual tokens from a single, fixed (typically penultimate) vision encoder layer. (b) **Dense Connector**: Aggregates features from multiple layers via concatenation or downsampling with a static fusion pattern. (c) **MMFuser**: Retrieves shallow-layer features using deep-layer queries through cross-attention to recover local details. (d) **TGIF (ours)**: Introduces a text-guided router that dynamically reweights visual features across layers based on the input query, enabling adaptive control over visual abstraction depth.

Single-layer projection. Most existing MLLMs, including LLaVA-style architectures, employ a simple MLP connector that projects visual tokens from a single, fixed layer of the vision encoder (typically the penultimate layer) into the LLM. While computationally efficient, this design exposes only late-stage semantic representations to the language model, limiting access to low-level and mid-level visual evidence that is often crucial for fine-grained grounding and hallucination verification.

Static multi-layer fusion. Recent works enrich visual representations by aggregating features from multiple encoder layers. DenseConnector concatenates or downsamples features from selected depths, while MMFuser retrieves shallow-layer details using deep-layer queries via cross-attention. Although these approaches improve visual richness, the fusion strategy remains *static*: the set of layers and their relative contributions are fixed once chosen and do not adapt to the input query.

Text-guided inter-layer fusion (TGIF). In contrast, TGIF treats the layers of a frozen vision encoder as a pool of specialized visual experts and dynamically selects among them conditioned on the input text. A lightweight router predicts a soft distribution over layers for each query, allowing the model to adaptively control which level of visual abstraction is exposed to the LLM. This enables conservative, low-level verification for hallucination-sensitive queries and higher-level semantic aggregation for descriptive or reasoning-oriented prompts.

Overall, this comparison highlights a key distinction between TGIF and prior fusion methods: rather than statically enriching visual representations, TGIF introduces query-dependent control over visual abstraction depth, directly addressing hallucination arising from mismatched visual evidence.

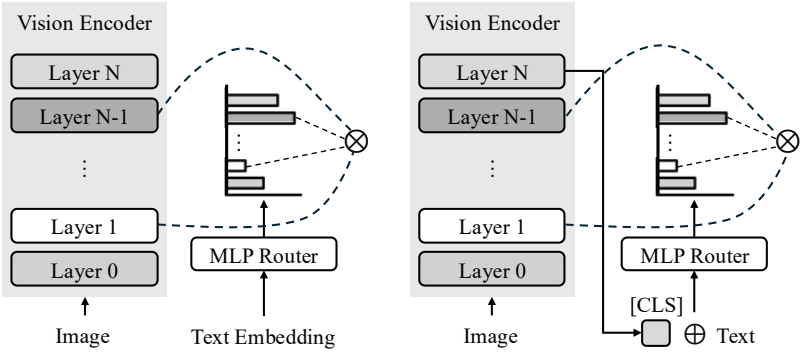

Figure 7: **Text-only vs. multimodal router in TGIF. Left:** The text-only router takes the question embedding $\mathbf{f}_{\text{text}}$ and outputs layer weights w_{layer} via an MLP and softmax. **Right:** The multimodal router concatenates a projected text embedding $\mathbf{f}_{\text{text}}\mathbf{W}_t$ and a projected global image embedding $\mathbf{f}_{\text{image}}\mathbf{W}_v$ (penultimate-layer [CLS]) to form $\mathbf{f}_{\text{multi}} \in \mathbb{R}^{2D_p}$ before predicting w_{layer}.

C Router Variants: Text-only vs. Multimodal

Figure 7 illustrates the difference between the two routing variants used in TGIF. The **text-only router** predicts layer weights solely from the question embedding, while the **multimodal router** additionally incorporates a global image representation (the penultimate-layer [CLS] token), enabling routing to depend on both query intent and coarse image content.

D Training Details

All experiments are conducted on a single node equipped with 8× NVIDIA H100 (80GB) GPUs. We utilize DeepSpeed ZeRO (Stage 2 for pretraining, Stage 3 for instruction tuning) to optimize memory efficiency. Unless otherwise specified, we employ Vicuna-7B-v1.5 as the language model and CLIP ViT-L/14@336px as the frozen vision encoder. Training is performed with `bf16` precision and `tf32` matrix multiplication acceleration.

D.1 Training Recipe

All models are trained following the two-stage training paradigm of LLaVA-1.5 Li et al. (2023a), consisting of **Stage 1: Feature Alignment Pretraining** and **Stage 2: Instruction Finetuning**. For a fair comparison, we adopt the same training datasets and data splits as LLaVA-1.5.

Stage 1: Feature Alignment. Following the LLaVA-1.5 protocol, we train only the multimodal projector (including the proposed TGIF router) for 1 epoch on the LLaVA pretraining dataset (558K image–text pairs). We utilize a learning rate of 1×10^{-3} with a cosine decay schedule and a warmup ratio of 0.03. The global batch size is set to 32 (gradient accumulation steps $= 1$), with a maximum sequence length of 2048. Weight decay is set to 0.

Stage 2: Instruction Tuning. In the second stage, we fine-tune the projector and the LLM on the LLaVA v1.5 mixture (665K samples) for 1 epoch. We reduce the learning rate to 2×10^{-5} while maintaining the cosine schedule, warmup ratio of 0.03, and weight decay of 0. The per-GPU batch size is adjusted to 16.

Throughout both stages, the vision encoder remains entirely frozen. We enable gradient checkpointing to conserve memory. All specific hyperparameters and script templates are provided in the attached code for reproducibility.

E Benchmark Descriptions

E.1 Hallucination Evaluation

POPE Yifan Li & Wen (2023). POPE assesses object hallucination via a binary verification task. For a given image, the model must answer Yes/No questions (e.g., "Is there a <object> in the image?"). Negative samples are generated using three strategies: (1) **Random** sampling, (2) **Popular** COCO categories, and (3) **Adversarial** co-occurring objects. This setup isolates visual grounding capabilities from captioning priors. We report Accuracy, Precision, Recall, F1-score, and the "Yes" response ratio.

HallusionBench Liu et al. (2024a) This benchmark evaluates visual factual grounding by presenting questions with fabricated or contradictory premises. Unlike standard VQA, HallusionBench penalizes models for failing to reject incorrect visual claims regarding object existence, attributes, and spatial relations. Following the official protocol, we employ a GPT-4 assisted evaluation metric. We report **aAcc** (All Accuracy) as the primary metric, alongside **qAcc** (Question-Pair Accuracy) and **fAcc** (Figure Accuracy).

E.2 OCR Evaluation

OCRBench Liu et al. (2024e). OCRBench is a comprehensive evaluation suite consisting of 1,000 manually verified QA pairs aggregated from 29 diverse datasets. It assesses five core capabilities: Text Recognition, Scene Text VQA, Document VQA, Key Information Extraction (KIE), and Handwritten Math Expression Recognition (HMER). Evaluation is performed via exact string matching.

TextVQA Singh et al. (2019a). TextVQA is a benchmark designed to evaluate a model's ability to read and reason about text embedded in natural images. It contains over 45K questions paired with images rich in scene text (e.g., signs, posters, menus), where answering explicitly requires OCR and text-based reasoning. The dataset features a highly diverse, open-ended answer space, making it a challenging testbed for fine-grained visual grounding and text-centric multimodal understanding.

E.3 General Reasoning and Overall Benchmarks.

ScienceQA Lu et al. (2022). ScienceQA is a large-scale multimodal multiple-choice benchmark designed to evaluate multi-hop scientific reasoning across vision and language. It contains over 21K questions drawn from elementary and high school science curricula, spanning natural, social, and language sciences. Nearly half of the questions include visual context, and most are annotated with grounded lectures and detailed explanations, enabling analysis of both answer accuracy and reasoning quality. ScienceQA is widely used to assess general multimodal reasoning and instruction-following capabilities beyond surface-level perception.

GQA Hudson & Manning (2019). GQA is a large-scale visual question answering benchmark designed to evaluate compositional and multi-step visual reasoning. It contains over 22M questions grounded in real-world images, each paired with a scene graph encoding objects, attributes, and relations. Questions are generated to require explicit reasoning over these structured representations, reducing language priors and dataset bias. In addition to accuracy, GQA provides fine-grained metrics to assess consistency and plausibility, making it a standard benchmark for general visual reasoning and scene understanding.

Table 9: **HallusionBench correctness leaderboard.** We report Question-pair Accuracy (*qAcc*), Figure Accuracy (*fAcc*), and All Accuracy (*aAcc*). Top-3 models under GPT-4–assisted evaluation are highlighted in **bold**.

Method	Params	Eval	qAcc ↑	fAcc ↑	aAcc ↑
GPT-4V 202 (2023)	–	Human	31.42	44.22	67.58
		GPT-4	**28.79**	**39.88**	**65.28**
LLaVA-1.5 Liu et al. (2023b)	13B	Human	9.45	25.43	47.12
		GPT-4	10.55	**24.86**	46.94
Claude 3 Team (2024)	–	GPT-4	**21.76**	**28.61**	**56.86**
Gemini Pro Vision Team (2023)	–	GPT-4	7.69	8.67	36.85
BLIP2-T5 Li et al. (2023c)	12.1B	GPT-4	15.16	20.52	48.09
Qwen-VL Bai et al. (2023)	9.6B	GPT-4	5.93	6.65	39.15
Open-Flamingo Alayrac et al. (2022)	9B	GPT-4	6.37	11.27	38.44
MiniGPT-5 Zheng et al. (2023)	8.2B	GPT-4	10.55	9.83	40.30
MiniGPT-4 Zhu et al. (2023b)	8.2B	GPT-4	8.79	10.12	35.78
InstructBLIP Dai et al. (2023a)	8.2B	GPT-4	9.45	10.11	45.26
BLIP-2 Li et al. (2023c)	8.2B	GPT-4	5.05	12.43	40.48
mPLUG-Owl v2 Ye et al. (2023c)	8.2B	GPT-4	13.85	19.94	47.30
mPLUG-Owl v1 Ye et al. (2023a)	7.2B	GPT-4	9.45	10.40	43.93
LRV-Instruction Liu et al. (2023a)	7.2B	GPT-4	8.79	13.01	42.78
LLaVA-1.5 + TGIF (Ours)	7B	GPT-4	**17.36**	23.70	**49.94**
GIT Wang et al. (2022)	0.8B	GPT-4	5.27	6.36	34.37
Random Chance	–	GPT-4	15.60	18.21	45.96

MMBench Liu et al. (2024d) MMBench is a large-scale, fine-grained benchmark designed to systematically evaluate the perceptual and reasoning abilities of vision–language models. It contains 2,974 multiple-choice questions spanning 20 ability dimensions, organized into a three-level taxonomy covering coarse perception, fine-grained perception, and multimodal reasoning. Instead of task-specific evaluation, MMBench assesses models across diverse capability dimensions, providing more diagnostic feedback. To ensure robust and reproducible evaluation, model outputs are matched to answer choices using an LLM-based evaluator, enabling reliable comparison across open-ended generation models.

F Additional Quantitative Results

F.1 HallusionBench Leaderboard

Table 9 presents the full HallusionBench correctness leaderboard. TGIF achieves competitive performance among 7B parameters models, particularly in the aAcc metric, surpassing the LLaVA-1.5 baseline and approaching the performance of proprietary models like Gemini Pro Vision.

Notably, despite utilizing a significantly smaller language backbone (7B parameters), TGIF (49.94%) outperforms several larger open-source baselines, including the 13B-parameter LLaVA-1.5 (46.94%) and the 12.1B-parameter BLIP2-T5 (48.09%) Liu et al. (2023b); Li et al. (2023c). It secures the third rank overall, trailing only the closed-source models GPT-4V and Claude 3 202 (2023); Team (2024). This result shows the efficiency of our text-guided fusion strategy: by dynamically routing to the most relevant visual features, TGIF extracts rich grounding signals from the frozen encoder, effectively mitigating hallucination even against models with nearly double the parameter count.

F.2 OCRBench Performance

We report detailed OCRBench scores in Table 10. TGIF demonstrates improvements in Scene Text VQA (VQA^S) and Document VQA (VQA^D), contributing to a higher final score compared to the LLaVA-1.5 baseline. This suggests that layer fusion effectively captures fine-grained textual details often lost in single-layer embeddings.

Table 10: **Detailed results on OCRBench.** Breakdown by sub-task: Text Recognition (Recog.), Scene Text VQA (VQA^S), Document VQA (VQA^D), Key Information Extraction (KIE), and Handwritten Math Expression Recognition (HMER). Best results are shown in **bold**.

Method	Recog.	VQA^S	VQA^D	KIE	HMER	Total
Gemini Pro Team (2023)	**215**	**174**	128	134	8	**659**
GPT-4V 202 (2023)	167	163	**146**	**160**	**9**	645
mPLUG-Owl2 Ye et al. (2023c)	153	153	41	19	0	366
LLaVAR Zhang et al. (2023)	186	122	25	13	0	346
LLaVA-1.5-13B Liu et al. (2023b)	176	129	19	7	0	331
LLaVA-1.5-7B Liu et al. (2023b)	160	117	15	5	0	297
+TGIF (Ours)	162	121	24	6	0	313
mPLUG-Owl Ye et al. (2023b)	172	104	18	3	0	297
InstructBLIP Dai et al. (2023b)	168	93	14	1	0	276
BLIP-2 Li et al. (2023c)	154	71	10	0	0	235
MiniGPT-4 v2 Zhu et al. (2023a)	124	29	4	0	0	157