MMEB-V3: Measuring the Performance Gaps of Omni-Modality Embedding Models

Haohang Huang[1]* **Xuan Lu**[1,2]* **Mingyi Su**[4] **Xuan Zhang**[5], **Ziyan Jiang**[6] **Ping Nie**[4], **Kai Zou**[7], **Tomas Pfister**[3], **Wenhu Chen**[4], **Wei Zhang**[1], **Xiaoyu Shen**[1]†, **Rui Meng**[3]
[1]Eastern Institute of Technology, Ningbo [2]Shanghai Jiao Tong University
[3]Google AI Research [4]University of Waterloo [5]NUS [6]UCSB [7]Netmind.ai
lux1997@sjtu.edu.cn haohang623@gmail.com xyshen@eitech.edu.cn

Code: https://github.com/TIGER-AI-Lab/VLM2Vec

Leaderboard: MMEB-V3

Abstract

Multimodal embedding models aim to map heterogeneous inputs, such as text, images, videos, and audio, into a shared semantic space. However, existing methods and benchmarks remain largely limited to partial modality coverage, making it difficult to systematically evaluate full-modality representation learning. In this work, we take a step toward the *full-modality* setting. We introduce **MMEB-V3**, a comprehensive benchmark that evaluates embeddings across text, image, video, audio, as well as agent-centric scenarios. To enable more fine-grained diagnosis, we further construct **OmniSET (Omni-modality Semantic Equivalence Tuples)**, where semantically equivalent instances are represented across modalities, allowing us to disentangle semantic similarity from modality effects. Through experiments on MMEB-V3, we conduct a systematic analysis of full-modality embeddings and identify three key findings: (1) models often fail to retrieve the intended target modality; (2) cross-modal retrieval is highly asymmetric and dominated by query-modality bias; and (3) instruction-induced shifts are either insufficient or misaligned with the target modality, and therefore do not reliably improve retrieval. These results indicate that current multimodal embeddings are not yet capable of reliably enforcing modality constraints specified by instructions, and consequently fail to exhibit consistent modality-aware retrieval behavior. We hope MMEB-V3 provides a useful benchmark for understanding and diagnosing these limitations, and for guiding future research on full-modality embeddings.

1 Introduction

Multimodal embeddings (Zhang et al., 2025a; Meng et al., 2025) are foundational to modern machine learning, mapping heterogeneous inputs—such as text, images, videos, and audio—into a unified, fixed-dimensional vector space. These representations power a vast ecosystem of applications, from semantic retrieval and recommendation systems to complex decision-making pipelines. As the field evolves, research has shifted from isolated unimodal encoders (Radford et al., 2021) toward *unified multimodal embedding models* that align diverse modalities. Such unified spaces are increasingly critical for emerging paradigms like multimodal Retrieval-Augmented Generation (RAG) and intelligent AI agents, which must retrieve and act upon information seamlessly across different sensory inputs.

Despite this progress, we identify a fundamental but underexplored challenge: **modality as an explicit instruction constraint**. Existing frameworks implicitly assume that modality is either perfectly aligned with semantics or acts as a passive attribute. In practice, however,

*Equal contribution.
†Corresponding author.

modality often serves as an explicit requirement specified by a user or system. For example, a user may issue queries such as "*find an audio clip of a cat meowing*" or "*retrieve a video showing a cat jumping*", where modality is a mandatory constraint rather than an optional attribute. This reveals a key limitation: *current embedding models often fail to interpret modality as an explicit instruction constraint, instead treating it as a byproduct of semantic similarity*, leading to semantically relevant but modality-mismatched retrieval results. This issue is particularly acute in agent-centric settings, where retrieval is a prerequisite for downstream actions such as tool invocation, GUI interaction, or memory access. Errors in modality understanding—such as retrieving semantically relevant but modality-mismatched results—can lead to catastrophic failures in the decision-making loop. However, current benchmarks such as MMEB (Meng et al., 2025) and UMR (Zhang et al., 2025a) provide limited support for evaluating this problem. They focus primarily on cross-modal alignment (e.g., text-to-image) and lack systematic evaluation of (1) full-modality coverage (especially audio and visual documents) and (2) instruction-following behavior in complex retrieval scenarios.

To systematically **measure and diagnose limitations in modality-aware, instruction-conditioned embedding behavior**, we introduce **MMEB-V3**, a comprehensive full-modality benchmark designed to evaluate embeddings under realistic, instruction-conditioned settings. MMEB-V3 extends prior work along four key dimensions: (1) **Audio Tasks**, expanding coverage to include audio classification, temporal grounding and cross-modal retrieval; (2) **Text Tasks**, incorporating instruction-following, reasoning, and multi-condition constraints; (3) **Agent Tasks**, evaluating embeddings in agent-centric scenarios such as tool retrieval, memory retrieval, and GUI control; and (4) **OmniSET (Omni-modality Semantic Equivalence Tuples)**, which organize semantically equivalent content across modalities into unified tuples, enabling controlled analysis that disentangles semantic content from modality effects. Through experiments on MMEB-V3, we conduct a systematic analysis of embedding behavior under instruction-conditioned settings. Our results show that while existing models perform well on standard semantic alignment, they struggle to enforce modality constraints specified by instructions. In many cases, retrieval is dominated by the original query modality rather than the instructed target modality, leading to systematic modality mismatches. These findings highlight a key limitation in current multimodal embeddings: modality constraints are not reliably enforced as part of the instruction.

Our main contributions are as follows:

- We introduce **MMEB-V3**, a comprehensive benchmark for evaluating full-modality embeddings under instruction-conditioned settings across diverse tasks.
- We provide a **systematic diagnostic analysis** of modality-aware behavior, enabled by **OmniSET (Omni-modality Semantic Equivalence Tuples)**, a controlled design that disentangles semantic content from modality effects.
- We conduct a **comprehensive analysis of modality-constrained retrieval**, showing that (i) models often fail to retrieve the intended target modality, (ii) retrieval is asymmetric and biased toward the query modality, and (iii) instruction-induced shifts are insufficient or misaligned with the target modality.

2 Related Work

Multimodal Embedding Benchmarks. The evolution of multimodal embedding benchmarks reflects a transition from coarse-grained cross-modal alignment to fine-grained, task-diverse evaluation. Early efforts primarily focused on image–text alignment, with benchmarks such as MSCOCO (Lin et al., 2015), Flickr30K (Plummer et al., 2015) and UMR (Zhang et al., 2025a) establishing standard protocols for retrieval and matching. While foundational, these benchmarks are largely constrained to static visual inputs and brief textual descriptions. Subsequent research has expanded this scope to specialized domains: ViDoRe-v2 (Macé et al., 2025) introduces document-level visual retrieval, while QVHighlights (Lei et al., 2021) and FineVideo (Farré et al., 2024) address temporal dynamics in video retrieval. Despite these advancements, a significant gap persists in evaluating instruction-driven, full-modality generalizability. While MTEB (Muennighoff et al., 2023) provides a comprehensive suite for unimodal text embeddings, it lacks multimodal support; conversely,

MMEB-V2 (Meng et al., 2025) and M-BEIR (Wei et al., 2023) integrate images, videos, and documents but omit other critical modalities such as audio. Furthermore, emerging agentic tasks—including tool retrieval (Lu et al., 2026a), GUI control (Zhang et al., 2025b) and memory retrieval (Zhao et al., 2026)—increasingly rely on embedding-based solutions, yet they remain excluded from existing comprehensive benchmarks. Existing multimodal embedding benchmarks mainly focus on semantic alignment across modalities (e.g., text–image retrieval), while more complex scenarios—such as instruction-conditioned retrieval, multi-condition reasoning, and agent tasks (e.g., tool, memory, and GUI retrieval)—remain under-explored.

Multimodal Embedding Models. Multimodal embedding models aim to learn shared representations for heterogeneous modalities (e.g., text, images, audio, and video) within a unified framework. Early approaches typically follow a dual-encoder paradigm, aligning modality-specific encoders in a common space, as exemplified by CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and their extensions to video and audio. Building on this foundation, recent work explores partial modality unification for embedding-based retrieval. Models such as GME (Zhang et al., 2025a), MM-Embed (Lin et al., 2025), VLM2Vec (Jiang et al., 2025), and VLM2Vec-V2 (Meng et al., 2025) unify subsets of modalities and achieve strong performance on their target benchmarks, often leveraging MLLM backbones or contrastive objectives. More recent efforts move toward broader modality coverage. Omni-Embed-Nemotron (Xu et al., 2025) supports retrieval across text, image, audio, and video, while OmniRet (Huynh et al., 2026) improves efficiency via multimodal token resampling and pooling strategies. WAVE (Tang et al., 2025) focuses on unified audio–visual embeddings, enabling cross-modal retrieval between audio and video. To complement these developments, we introduce MMEB-V3, which extends prior benchmarks with audio modalities, complex text retrieval, and agent tasks, enabling more comprehensive evaluation under realistic, instruction-conditioned settings.

3 MMEB-V3: A Unified Evaluation Framework for Omni-Modality Embeddings

3.1 Modality Coverage and Task Diversity

MMEB-V3 extends MMEB-V2 by adding **111** new tasks (see Table 1). It provides an evaluation framework covering four modalities—text, image, video, and audio—encompassing a total of **190** tasks. Beyond expanding modality coverage, MMEB-V3 organizes tasks across diverse cross-modal directions (e.g., A2I, T2A, A2V), allowing evaluation under a range of modality interactions. On top of this modality space, it includes multiple task types, such as audio classification, cross-modal retrieval, multi-condition text retrieval, and agent tasks. This design reflects both modality diversity and task diversity, with tasks evaluated under specific modality configurations rather than in isolation.

In addition to standard evaluation, MMEB-V3 includes a diagnostic component, OmniSET (Omni-modality Semantic Equivalence Tuples), which groups semantically equivalent instances across modalities to facilitate controlled comparisons. OmniSET is not used for leaderboard evaluation, but provides a setting to analyze the relationship between semantic content and modality effects. Figure 1 provides an overview of MMEB-V3. Figure 2a further illustrates the distribution of tasks across modalities and task categories.

Audio Tasks. MMEB-V3 incorporates diverse audio tasks, including classification (e.g., ESC-50 (Piczak, 2015), UrbanSound8K (Salamon et al., 2014), NSynth (Engel et al., 2017)), cross-modal retrieval (e.g., Clotho (Drossos et al., 2020), SoundDescs (Koepke et al., 2023), AVE (Tian et al., 2018), SpeechCOCO (Havard et al., 2017)), and temporal grounding (TUT Sound Events (Mesaros et al., 2016)). For TUT Sound Events, we construct two tasks with different difficulty levels and use their average score as the final temporal grounding metric. The audio tasks evaluate both intra-modal understanding and alignment between audio and other modalities, providing a basis to assess acoustic-semantic representations.

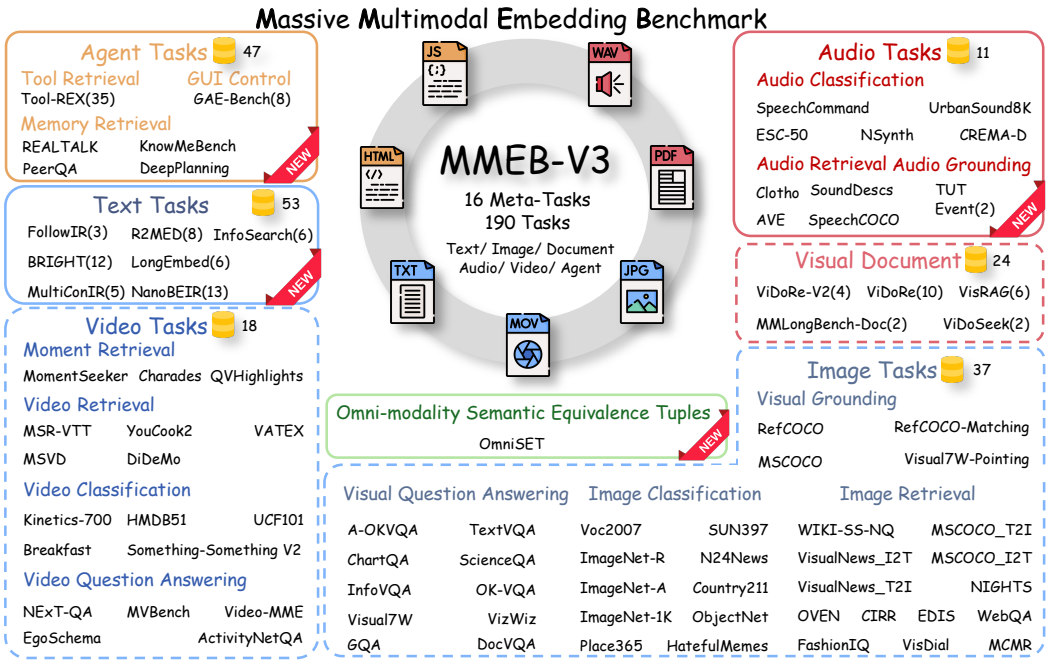


Figure 1: MMEB-V3 Overview: New Additions—Agent Tasks, Complex Text Retrieval, Audio Tasks and Equivalence Tuples; Built upon Image, Video, and VisDoc from MMEB-V2

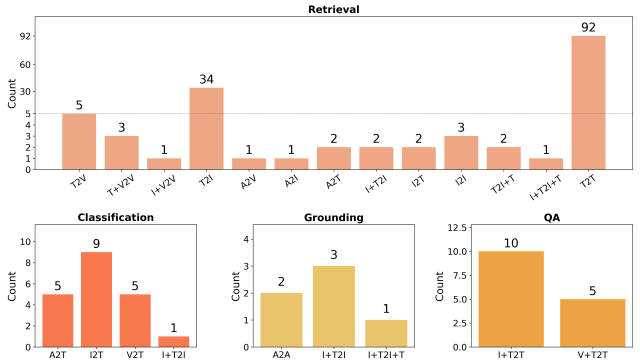
Text Tasks. Beyond standard semantic retrieval, MMEB-V3 includes text scenarios that involve instruction following (FollowIR (Weller et al., 2025), InfoSearch (Zhou et al., 2025)), reasoning (BRIGHT (SU et al., 2025), R2MD (Li et al., 2025)), long-context understanding (LongEmb (Zhu et al., 2024)), and multi-condition matching (MultiConIR (Lu et al., 2025)). A compact general retrieval benchmark (nanoBEIR (Thakur et al., 2021)) is also included to provide coverage of general retrieval settings. These datasets are composed of multiple tasks, resulting in a total of 53 tasks. The final Text score is computed as the average of NDCG@5 across these tasks.

Agent Tasks. MMEB-V3 includes agent tasks, such as tool retrieval (Tool-REX (Lu et al., 2026a)), GUI trajectory retrieval (GAE-Bench (Zhang et al., 2025b)), and memory retrieval (KnowMeBench (Wu et al., 2026), REALTALK (Lee et al., 2025), PeerQA (Baumgärtner et al., 2025), and DeepPlanning (Zhang et al., 2026)). These tasks involve selecting tools or actions, or retrieving relevant information from past memory based on multimodal inputs and structured instructions. These datasets are composed of multiple tasks, resulting in a total of 47 tasks. The final Agent score is computed as the average of Hit@1 across these tasks.

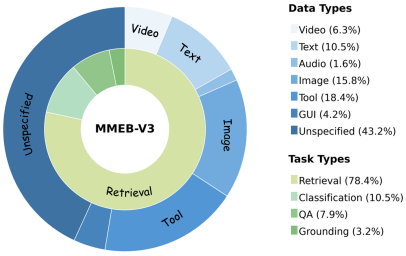
Cross-Modal Equivalence Modeling. We include a diagnostic component, OmniSET (Omni-modality Semantic Equivalence Tuples), which groups semantically equivalent instances across modalities into tuples $\{x^T, x^I, x^V, x^A\}$. OmniSET is not used for leaderboard evaluation, but provides a setting for controlled comparisons between semantic content and modality effects. It enables analysis of how models handle modality as an explicit constraint under instruction-conditioned retrieval.

3.2 Instruction Diversity across Data Types and Tasks

We describe instruction diversity in MMEB-V3 from two perspectives: **target-type constraints** and **task-type diversity**. First, **Data-type constraints** refer to the expected type of returned data, including modalities such as **text**, **image**, **audio**, and **video**, as well as structured data such as **tools** and **GUI elements**. This dimension captures whether and how instructions constrain the modality or representation of the data. Second, **Task-type di-**



(a) Distribution of modality combinations across task types in MMEB-V3. This illustrates the comprehensive modality coverage and the diversity of cross-modal interactions and task types.



(b) Distribution of instruction patterns, including data-type constraints and task-type diversity. This highlights that instructions vary in their data requirements as well as the diversity of underlying task types.

Figure 2: Diversity of modalities, tasks, and instruction patterns in MMEB-V3 **(190 tasks)**. (a) Distribution of modality combinations across task types, demonstrating comprehensive modality coverage and rich cross-modal interactions across diverse tasks. (b) Distribution of instruction patterns, including data-type constraints and task-type diversity, showing that instructions are not homogeneous but vary in both data requirements and associated task contexts.

versity corresponds to the range of operations associated with different tasks. In MMEB-V3, these mainly include **classification**, **retrieval**, **grounding**, and **QA**, covering behaviors ranging from label prediction and candidate selection to localization and question answering.

Together, these perspectives offer a way to describe instruction diversity in MMEB-V3. This formulation shows that tasks vary in both *what* type of output is required and *what* operation is involved. Figure 2b visualizes the distribution of these patterns across MMEB-V3.

4 Experiments

4.1 Experimental Settings

Metrics. MMEB-V3 uses task-appropriate metrics. We adopt **Hit@1** for audio, image, video, and agent tasks, where the number of relevant targets is typically small and the evaluation focuses on identifying a highly relevant match. We use **NDCG@5** for text and VisDoc tasks, which involve multiple relevant candidates and require fine-grained ranking. This design aligns evaluation metrics with task characteristics.

Baselines. We evaluate against two groups of baselines, including omni-modal embedding models (Omni-Embed-Nemotron (Xu et al., 2025) and WAVE (Tang et al., 2025)) and vision-language embedding models (Qwen3-VL-Embedding (Li et al., 2026), VLM2Vec-V2.0 (Meng et al., 2025), VLM2Vec (Jiang et al., 2024), E5-Omni (Chen et al., 2026), LCO-Embedding-Omni (Xiao et al., 2026) and GME (Zhang et al., 2025a)).

4.2 Main Results

Table 3 presents a comparison of representative multimodal embedding models on MMEB-V3, covering text, image, video, audio, visual document, and agent tasks. Full results across all tasks are provided in Appendix A.3. A general pattern can be observed: no single model achieves optimal performance across all task categories. Models that perform well on certain modalities or tasks often exhibit weaker performance on others, suggesting trade-offs in current embedding approaches.

Task	Task type	Query MOD	Target MOD	Domain	#Query	#Candidates
Audio Tasks (11 tasks)						
ESC-50	Audio Classification	A	T	Environmental sounds	2K	50
UrbanSound8K	Audio Classification	A	T	Environmental sounds	1.7K	10
NSynth	Audio Classification	A	T	Music	1K	10
Speech Commands	Audio Classification	A	T	Human Voice	1K	36
CREMAD	Audio Classification	A	T	Human Voice	7.4K	6
Clotho	Audio Retrieval	T	A	Open	5.2K	1K
Sound-Descs	Audio Retrieval	T	A	Open	1K	4.9K
AVE	Audio Retrieval	A	V	Acoustic Events	402	402
SpeechCOCO	Audio Retrieval	A	I	Open	1K	10K
TUT Sound Events (2)	Audio Temporal Grounding	A	A	Acoustic Events	659	$51 \sim 106$
Text Tasks (53 tasks)						
FollowIR (3)	Instruction-following	T	T	Open	104	98K
InfoSearch (6)	Instruction-following	T	T	Open	1.6K	6.3K
BRIGHT (12)	Reasoning Retrieval	T	T	Open	1.4K	1.3M
R2MD (8)	Reasoning Retrieval	T	T	Medical	876	357K
LongEmb (6)	Long-context Retrieval	T	T	Open	13K	2.7K
MultiConIR (5)	Multi-condition Retrieval	T	T	Open	9K	25K
nanoBEIR (13)	General Text Retrieval	T	T	Open	649	56K
Agent Tasks (47 tasks)						
Tool-REX (35)	Tool Retrieval	T	T	Tool Use	7.9K	44K
GAE-Bench (8)	GUI Control	I/T	I/T	GUI	8.2K	78K
KnowMeBench	Agent Memory Retrieval	T	T	Episodic Memory	2K	27K
REALTALK	Agent Memory Retrieval	T	T	Dialogue Memory	679	8.9K
PeerQA	Agent Memory Retrieval	T	T	Semantic Memory	136	18.6K
DeepPlanning	Agent Memory Retrieval	T	T	Procedural Memory	120	19.8K
Omni-modality Semantic Equivalence Tuples						
OmniSET	Omni-modality Semantic Equivalence Tuples	T/I/V/A	T/I/V/A	Open	1.2K	8.2K

Table 1: The statistics of MMEB-V3 are summarized below. Compared to MMEB-V2, MMEB-V3 introduces **111 new tasks** across three major categories: audio tasks, agent tasks, and complex text retrieval tasks, resulting in a total of **190 tasks**. We report the detailed statistics of these newly added tasks. We also include statistics of OmniSET as a diagnostic component for controlled analysis. OmniSET is used for comparative analysis of modality effects and is not part of the leaderboard evaluation. Modalities (MOD): T (Text), I (Image), V (Video), and A (Audio).

For example, models such as Qwen3-VL-Embedding achieve strong performance on text, image, and video tasks, but lack native audio capability, which may limit their applicability in settings involving audio. Fully multimodal models (e.g., E5-Omni) show relatively stable performance across modalities, yet do not consistently outperform specialized models within specific task categories. To further understand these trade-offs, Table 2 provides a fine-grained breakdown of newly introduced audio, text, and agent tasks. We observe that several task types are relatively challenging across models, including audio retrieval, as well as complex text and agent tasks involving reasoning, multi-condition constraints, and long-context understanding. This may indicate limitations in capturing fine-grained semantic structure and implicit constraints.

Overall, these findings suggest that the observed trade-offs are associated with both modality differences and task complexity, as well as instruction requirements, pointing to the importance of developing more unified and instruction-aware embedding models.

5 Analysis

We analyze three representative embedding models on the OmniSET dataset: `Omni-Embed-Nemotron-3B`, `WAVE`, and `Qwen3-VL-Embedding-8B`. OmniSET provides semantically aligned instances across modalities, enabling controlled analysis of modality effects and instruction-conditioned retrieval.

Model	Audio				Text						Agent				All*	All
	CLS	RET	TG	Overall	RR	IF	LC	MC	GR	Overall	Tool	GUI	Memory	Overall		
# of Datasets →	5	4	2	11	20	9	6	5	13	53	35	8	4	47	100	111
Qwen3-VL-Embedding (2B)	-	-	-	-	16.6	40.6	53.1	61.2	58.0	39.2	**42.6**	30.4	28.4	**39.3**	39.2	35.4
Qwen3-VL-Embedding (8B)	-	-	-	-	**18.2**	44.8	**58.0**	61.2	61.2	**42.5**	41.3	33.5	22.8	38.4	**40.6**	36.5
VLM2Vec-Qwen2VL (7B)	-	-	-	-	7.2	28.1	5.9	40.9	41.7	22.2	19.8	21.4	15.9	19.7	21.0	19.0
VLM2Vec-V2.0 (2B)	-	-	-	-	7.8	29.2	11.5	50.3	41.2	24.5	27.6	**36.2**	23.3	28.7	26.5	23.9
GME (7B)	-	-	-	-	12.5	**52.4**	17.8	59.0	**62.5**	37.1	39.0	30.0	17.1	35.6	36.4	32.8
WAVE (7B)	52.3	12.9	21.5	31.8	5.9	31.3	2.6	22.9	18.6	13.7	11.9	11.8	5.7	11.3	14.3	14.3
Omni-Embed-Nemotron (3B)	53.7	18.8	29.1	36.5	17.2	42.8	40.0	**69.7**	56.9	39.2	38.1	32.0	**32.2**	36.5	37.6	**37.6**
LCO-Embedding-Omni (7B)	**64.3**	19.9	**36.8**	**43.2**	13.6	52.0	12.6	61.4	52.1	32.4	29.0	25.0	23.0	27.8	30.9	30.9
E5-Omni (3B)	41.4	22.0	21.6	30.8	11.5	44.4	8.3	60.7	34.4	26.7	37.7	35.6	32.3	36.9	31.2	31.2
E5-Omni (7B)	63.0	**23.5**	32.2	43.0	11.1	42.1	9.7	56.9	35.9	26.9	37.4	38.0	27.3	36.7	32.6	32.6

Table 2: Performance comparison across Audio, Text, and Agent tasks. CLS: classification; RET: retrieval; TG: temporal grounding; RR: reasoning retrieval; IF: instruction following; LC: long-context retrieval; MC: multi-condition retrieval; GR: general retrieval; Memory: agent memory retrieval. Tool: tool retrieval; GUI: GUI control. * *indicates that the overall score is averaged over available tasks only; the final All column averages over all modalities, treating missing ones as 0.*

Model	Image	Video	VisDoc	Audio	Text	Agent	All*	All
# of Datasets →	37	18	24	11	53	47	179	190
Qwen3-VL-Embedding (2B)	69.5	55.9	70.6	-	39.2	**39.3**	51.4	48.4
Qwen3-VL-Embedding (8B)	**72.1**	**58.6**	70.9	-	**42.4**	38.4	**53.0**	**49.9**
VLM2Vec-Qwen2VL (7B)	63.6	33.8	32.6	-	22.2	19.7	32.7	30.8
VLM2Vec-V2.0 (2B)	63.3	34.7	68.6	-	24.5	28.7	40.6	38.2
GME (7B)	55.2	38.4	75.2	-	37.1	35.6	45.7	43.0
WAVE (7B)	41.5	43.1	42.8	31.8	13.7	11.3	26.3	26.3
Omni-Embed-Nemotron (3B)	43.4	41.4	71.1	36.5	39.2	36.5	43.5	43.5
LCO-Embedding-Omni (7B)	42.2	48.4	68.7	**43.2**	32.4	27.8	39.9	39.9
E5-Omni (3B)	63.5	48.9	73.5	30.8	26.7	36.9	44.6	44.6
E5-Omni (7B)	70.5	50.8	**75.4**	43.0	26.9	36.7	47.1	47.1

Table 3: Performance comparison across Image, Video, VisDoc, Audio, Text, and Agent tasks. * *indicates that the overall score is averaged over available tasks only; the final All column averages over all modalities, treating missing ones as 0.*

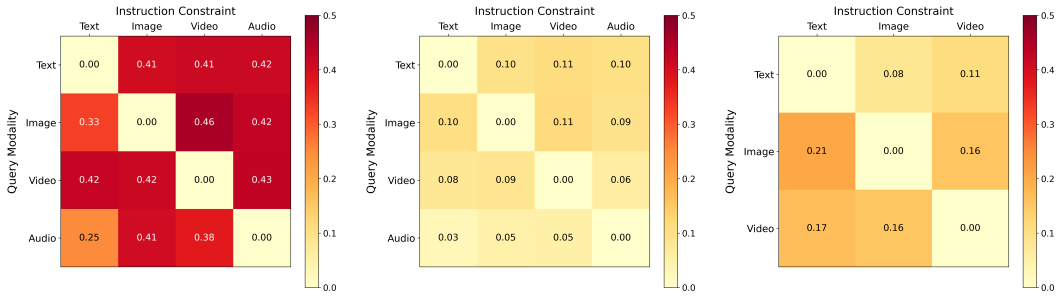
5.1 Explicit Modality Instructions Often Fail in Cross-Modal Retrieval

Key finding: Explicit modality instructions do not reliably lead to correct target-modality retrieval, and cross-modal behavior exhibits both asymmetry and modality bias. We first examine retrieval performance under explicit modality constraints. As shown in Table 4, most cross-modal directions remain challenging across all three models. In particular, Hit@1 is close to zero in many cases (e.g., T→I, T→A, V→T), and even MRR remains low, indicating that the instructed target modality is often not retrieved. This suggests that modality-specific instructions alone are insufficient to consistently guide retrieval toward the desired modality. A small number of directions achieve relatively high performance, such as I→V and A→T. However, these cases should be interpreted with caution. In OmniSET, videos are generated from images and audio from text, which introduces stronger intrinsic similarity for certain modality pairs. As a result, performance in these directions may partially reflect dataset construction effects rather than purely instruction-following capability. We provide a detailed discussion of this potential bias in Appendix A.2.2. Beyond overall performance, two consistent patterns emerge.

(1) Cross-Modal Asymmetry. Cross-modal retrieval is highly directional. Strong performance in one direction does not imply comparable performance in the reverse. For example, I→V achieves near-perfect scores (Hit@1 = 100.0 for Nemotron), while V→I remains low

Model	Metric	T2I	T2V	T2A	I2T	I2V	I2A	V2T	V2I	V2A	A2T	A2I	A2V
Omni-Embed Nemotron	Hit@1	0.0	1.5	0.0	0.0	100.0	0.0	0.0	5.0	0.0	100.0	0.0	0.0
	MRR	4.1	16.4	6.4	10.4	100.0	1.7	2.4	19.4	4.0	100.0	4.5	29.8
	Top-10 DM	T(84.9%)	T(84.7%)	T(75.6%)	V(95.9%)	V(95.8%)	V(95.5%)	I(99.5%)	I(99.5%)	I(99.5%)	T(71.3%)	T(62.8%)	T(63.6%)
WAVE	Hit@1	0.0	66.3	0.0	0.0	90.6	0.0	0.0	0.0	0.0	0.0	0.0	63.4
	MRR	2.6	77.3	2.7	2.5	94.4	2.7	2.5	2.6	2.7	2.5	2.6	75.5
	Top-10 DM	V(99.9%)	V(99.9%)	V(99.9%)	V(99.9%)	V(99.9%)	V(99.9%)	V(99.9%)	V(99.9%)	V(99.9%)	V(99.9%)	V(99.9%)	V(99.9%)
Qwen3-VL Embedding	Hit@1	0.0	0.0	–	0.0	100.0	–	0.0	2.5	–	–	–	–
	MRR	6.9	2.9	–	6.6	100.0	–	4.2	15.9	–	–	–	–
	Top-10 DM	T(81.7%)	T(84.0%)	–	V(97.8%)	V(97.7%)	–	I(99.9%)	I(99.9%)	–	–	–	–

Table 4: Cross-modal retrieval performance across three models. DM denotes the dominant modality among the top-10 retrieved results.



(a) `Omni-Embed-Nemotron`. (b) `WAVE`. (c) `Qwen3-VL-Embedding-8B`.

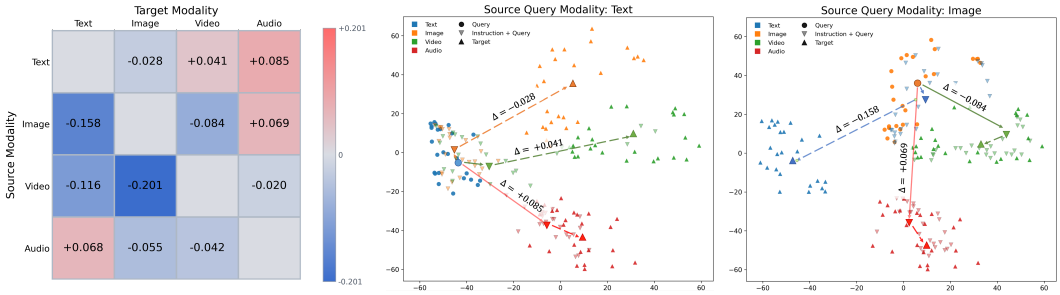
Figure 3: **Model sensitivity to modality-constrained instructions**, measured by the mean cosine distance between raw and instruction-augmented queries. Larger values indicate greater embedding shifts after instruction augmentation.

(Hit@1 = 5.0). Similar asymmetry appears in other modality pairs (e.g., A→T vs. T→A), indicating that cross-modal relations are not bidirectionally aligned in the embedding space. **(2) Modality bias.** Retrieval results are dominated by modality proximity rather than target constraints. The Top-10 dominant modality (DM) is strongly correlated with the query modality across all models. For instance, Omni-Embed-Nemotron retrieves predominantly text for text queries (e.g., T2I: 84.9%), while WAVE consistently retrieves video regardless of the target modality (99.9% across all directions). Qwen3-VL shows a similar pattern, where retrieved modalities align more with the query modality than with the instructed target.

5.2 Model Sensitivity to Modality-Constrained Instructions

Key finding: Models vary substantially in their sensitivity to modality-constrained instructions, ranging from strong responsiveness to near invariance. To better understand the retrieval failures, we examine how strongly each model responds to modality-constrained instructions at the representation level. We quantify this effect by measuring the cosine distance between the raw query embedding and its instruction-augmented counterpart, which captures the *magnitude* of the instruction-induced shift.

Figure 3 shows clear differences across models. `Omni-Embed-Nemotron` is highly sensitive to modality-constrained instructions, with an average shift of about 0.4 cosine distance; for most cross-modal directions, the shift exceeds 0.3. In contrast, `WAVE` is much less responsive, with an average change of only 0.08, indicating that modality-constrained instructions have limited effect on its query representations. `Qwen3-VL-Embedding-8B` falls between these extremes: across text–image–video tasks, its average shift is about 0.15, suggesting moderate responsiveness. These differences may be related to training objectives. Although Nemotron and WAVE both build on the Qwen2.5-Omni family, their responses to modality-constrained instructions differ substantially. While the instruction-related training details of Nemotron



(a) `Omni-Embed-Nemotron` (distance change heatmap).

(b) Text query (t-SNE visualization).

(c) Image query (t-SNE visualization).

Figure 4: **Instruction-induced changes in query–target distance and embedding shifts.** (a) Change in cosine distance to the target modality after instruction augmentation, measured relative to the raw query. (b,c) t-SNE visualizations showing how instruction-augmented queries move in the embedding space for text and image queries. In (b,c), raw queries are shown as circles (●), instruction-augmented queries as downward triangles (▼), and target instances as upward triangles (▲).

are not fully specified, its strong sensitivity and strong full-modality performance suggest that it likely benefited from more extensive instruction-related multimodal training. By contrast, WAVE focuses primarily on unified representation learning for visual and audio tasks, with training centered on multimodal retrieval and video QA rather than instruction-following. This difference in training emphasis may help explain its weaker response to modality-constrained instructions. Overall, sensitivity to modality instructions is clearly not uniform across models. Some models barely react to such constraints, whereas others exhibit substantial embedding shifts. However, as we show next, stronger sensitivity does not necessarily imply better alignment with the target modality.

5.3 Instruction-Induced Shifts Do Not Consistently Move Queries Toward the Target

Key finding: Instruction-induced shifts do not consistently reduce the distance to the target modality, and often move queries farther away. We next analyze the *direction* of instruction-induced shifts, i.e., whether instruction augmentation moves queries closer to the intended target modality. We compare the distance from the raw query to the target with that from the instruction-augmented query to the same target. We focus on `Omni-Embed-Nemotron`, which exhibits the strongest sensitivity to modality-constrained instructions; visualizations for other models are provided in Appendix A.4. As shown in Figure 4a, instruction-induced shifts do not consistently improve alignment. For Nemotron, only a few directions (e.g., T→V, T→A, A→T, I→A) show slight improvements, all below 0.09, while most directions instead increase the distance to the target modality. Moreover, this behavior is *asymmetric* across directions. For instance, T→V exhibits a small improvement (+0.041), whereas the reverse direction V→T shows a much larger degradation (-0.158). This asymmetry suggests that instruction-induced shifts do not establish a consistent bidirectional alignment between modality pairs, but instead interact unevenly with the underlying embedding geometry. The t-SNE visualizations in Figures 4b and 4c provide qualitative evidence. Although instruction augmentation induces noticeable movement in the embedding space, these shifts are not consistently oriented toward the target modality and can deviate toward other modality clusters. For example, image queries augmented with a text-target instruction are often observed to move closer to the video cluster rather than the text cluster. A similar trend is observed for `Qwen3-VL-Embedding-8B` (see Appendix A.4), where instruction-induced shifts yield only limited improvements in a small number of directions, while most changes do not reduce the query–target distance. Overall, instruction-induced shifts do not reliably translate into target-oriented movement. Even when models are sensitive to modality instructions, the induced changes are not consistently aligned with the target modality, limiting their effectiveness for cross-modal retrieval.

6 Conclusion

In this work, we advance the evaluation of multimodal embeddings toward the *full-modality* setting. We introduce MMEB-V3, a comprehensive benchmark that systematically evaluates embeddings across text, image, video, audio, and agent-centric scenarios. To enable controlled analysis, we further construct Omni-modality Semantic Equivalence Tuples (OmniSET), which isolate modality effects under consistent semantics and enable rigorous evaluation of cross-modal behavior under instruction constraints. Based on this framework, we conduct a systematic analysis of modality-constrained retrieval and identify several key limitations. We find that models often fail to retrieve the intended target modality, cross-modal retrieval is highly asymmetric and exhibits strong directional biases, and instruction-conditioned signals do not reliably guide retrieval toward the desired modality. These findings suggest that current multimodal embeddings remain limited in supporting reliable modality-aware retrieval under instruction constraints. We hope MMEB-V3 provides a useful benchmark for diagnosing these limitations and for guiding future research toward more controllable full-modality embedding models, particularly in emerging agent-centric applications.

Ethics Statement

We develop and evaluate general-purpose multimodal retrieval and embedding systems using publicly available datasets. No private or personally identifiable data is used. While our benchmarks may inherit biases from underlying corpora, we do not explicitly address fairness, and such biases may affect downstream applications. The proposed methods are intended for beneficial uses such as information access, but may have dual-use risks if misapplied. We encourage responsible use, careful dataset curation, and transparency through the release of code and evaluation protocols.

References

Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. Peerqa: A scientific question answering dataset from peer reviews, 2025. URL https://arxiv.org/abs/2502.13668.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Haonan Chen, Sicheng Gao, Radu Timofte, Tetsuya Sakai, and Zhicheng Dou. e5-omni: Explicit cross-modal alignment for omni-modal embeddings. *arXiv preprint arXiv:2601.03666*, 2026.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

William Havard, Laurent Besacier, and Olivier Rosec. Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set. *arXiv preprint arXiv:1707.08435*, 2017.

Chuong Huynh, Manh Luong, and Abhinav Shrivastava. Omniret: Efficient and high-fidelity omni modality retrieval, 2026. URL https://arxiv.org/abs/2603.02098.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation

learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks, 2025. URL https://arxiv.org/abs/2410.05160.

A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685, 2023. ISSN 1941-0077. doi: 10.1109/tmm.2022.3149712. URL http://dx.doi.org/10.1109/TMM.2022.3149712.

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. Realtalk: A 21-day real-world dataset for long-term conversation, 2025. URL https://arxiv.org/abs/2502.13270.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries, 2021. URL https://arxiv.org/abs/2107.09609.

Lei Li, Xiao Zhou, and Zheng Liu. R2med: A benchmark for reasoning-driven medical retrieval, 2025. URL https://arxiv.org/abs/2505.14558.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Chen Keqin, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms, 2025. URL https://arxiv.org/abs/2411.02571.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Xuan Lu, Sifan Liu, Bochao Yin, Yongqi Li, Xinghao Chen, Hui Su, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. MultiConIR: Towards multi-condition information retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13471–13494, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.726. URL https://aclanthology.org/2025.findings-emnlp.726/.

Xuan Lu, Haohang Huang, Rui Meng, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. Tools are under-documented: Simple document expansion boosts tool retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=g9D9MgG7iW.

Xuan Lu, Haohang Huang, Rui Meng, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. Rethinking reasoning in document ranking: Why chain-of-thought falls short. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=txmqENuRcc.

Xuan Lu, Kangle Li, Haohang Huang, Rui Meng, Wenjun Zeng, and Xiaoyu Shen. Beyond global similarity: Towards fine-grained, multi-condition multimodal retrieval, 2026c. URL https://arxiv.org/abs/2603.01082.

Quentin Macé, António Loison, and Manuel Faysse. Vidore benchmark v2: Raising the bar for visual retrieval, 2025. URL https://arxiv.org/abs/2505.17166.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents, 2025. URL https://arxiv.org/abs/2507.04590.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *2016 24th European signal processing conference (EUSIPCO)*, pp. 1128–1132. IEEE, 2016.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark, 2023. URL https://arxiv.org/abs/2210.07316.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ykuc5q381b.

Changli Tang, Qinfan Xiao, Ke Mei, Tianyi Wang, Fengyun Rao, and Chao Zhang. Wave: Learning unified & versatile audio-visual embeddings with multimodal llm, 2025. URL https://arxiv.org/abs/2509.21990.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021. URL https://arxiv.org/abs/2104.08663.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos, 2018. URL https://arxiv.org/abs/1803.08842.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition, 2018. URL https://arxiv.org/abs/1804.03209.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. Followir: Evaluating and teaching information retrieval models to follow instructions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11926–11942, 2025.

Tingyu Wu, Zhisheng Chen, Ziyan Weng, Shuhe Wang, Chenglong Li, Shuo Zhang, Sen Hu, Silin Wu, Qizhen Lan, Huacan Wang, and Ronghao Chen. Knowme-bench: Benchmarking person understanding for lifelong digital companions, 2026. URL https://arxiv.org/abs/2601.04745.

Chenghao Xiao, Hou Pong Ken Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. *Advances in Neural Information Processing Systems*, 38:158370–158401, 2026.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video, 2025. URL https://arxiv.org/abs/2510.03458.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms, 2025a. URL https://arxiv.org/abs/2412.16855.

Xuan Zhang, Ziyan Jiang, Rui Meng, Yifei Leng, Zhenbang Xiao, Zora Zhiruo Wang, Yanyi Shang, and Dehan Kong. Universal retrieval for multimodal trajectory modeling. In *ICML 2025 Workshop on Computer Use Agents*, 2025b.

Yinger Zhang, Shutong Jiang, Renhao Li, Jianhong Tu, Yang Su, Lianghao Deng, Xudong Guo, Chenxu Lv, and Junyang Lin. Deepplanning: Benchmarking long-horizon agentic planning with verifiable constraints, 2026. URL https://arxiv.org/abs/2601.18137.

Xinping Zhao, Xinshuo Hu, Jiaxin Xu, Danyu Tang, Xin Zhang, Mengjia Zhou, Yan Zhong, Yao Zhou, Zifei Shan, Meishan Zhang, Baotian Hu, and Min Zhang. Lmeb: Long-horizon memory embedding benchmark, 2026. URL https://arxiv.org/abs/2603.12572.

Jianqun Zhou, Yuanlei Zheng, Wei Chen, Qianqian Zheng, Shang Zeyuan, Wei Zhang, Rui Meng, and Xiaoyu Shen. Beyond content relevance: Evaluating instruction following in retrieval models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OlRjxSuSwl.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Longembed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 802–816, 2024.

A Appendix

A.1 Details of Baseline Models

Omni-Embed-Nemotron (Xu et al., 2025) is a unified multimodal embedding model designed for retrieval over text, images, audio, and video. Built on the Thinker component of Qwen2.5-Omni-3B, it encodes different modalities into a shared embedding space and is trained with a bi-encoder contrastive objective for multimodal retrieval and RAG scenarios.

WAVE (Tang et al., 2025) is a unified audio-visual embedding model for cross-modal retrieval and multimodal understanding. It learns shared audio-visual representations through joint multimodal training and hierarchical feature fusion. Combined with unified embedding learning, it achieves strong performance on audio-visual retrieval and multimodal question answering tasks.

Qwen3-VL-Embedding (Li et al., 2026) is a multimodal embedding model built on the Qwen3-VL foundation model for cross-modal retrieval and understanding. It supports text, images, screenshots, videos, and mixed-modal inputs within a unified framework, learning unified representations across modalities and languages. Using dense embedding representations, it enables high-quality multimodal retrieval and clustering.

VLM2Vec-V2.0 (Meng et al., 2025) is a multimodal embedding model that extends unified representation learning to diverse visual modalities. It supports images, videos, and visual documents within a shared embedding space, and employs unified training across modalities to learn consistent representations. Using contrastive embedding learning, it enables effective retrieval in diverse multimodal settings.

VLM2Vec (Jiang et al., 2024) converts an instruction-tuned vision-language model into a unified multimodal embedding model for retrieval. It reformulates multimodal inputs into embedding-oriented representations under a unified framework, enabling consistent encoding across modalities. Using contrastive embedding learning, it achieves strong performance across diverse multimodal retrieval tasks.

GME (Zhang et al., 2025a) is a unified multimodal embedding model based on Qwen2-VL. It supports three input types—text, image, and image-text pairs—and maps both single-modal and combined-modal inputs into universal vector representations, enabling versatile Any2Any retrieval scenarios such as text-to-image, image-to-image, and multimodal search.

A.2 Details of Benchmark Construction

MMEB-V3 represents a significant advancement over its predecessor, MMEB-V2, by establishing *a comprehensive unified embedding evaluation framework* that encompasses text, image, video, and audio modalities. While previous iterations focused predominantly on static images and eventually expanded to videos and visual documents, MMEB-V3 fills critical gaps in the embedding landscape by integrating comprehensive audio support, complex text reasoning, and specialized agentic capabilities. In addition, we further enrich the image domain by introducing a multi-condition, multimodal retrieval dataset, **MCMR** (Lu et al., 2026c), which evaluates fine-grained cross-modal matching under multiple constraints. By consolidating 190 heterogeneous tasks into a standardized ranking-based evaluation, MMEB-V3 provides a rigorous testbed for developing general-purpose, omni-modality embeddings capable of instruction-following and cross-modal semantic alignment.

Audio Tasks. To address the historical scarcity of audio-focused evaluation in unified models, MMEB-V3 introduces a suite of audio tasks covering classification, cross-modal retrieval, and temporal grounding. For several large-scale audio datasets, including NSynth, SpeechCommands, SoundDescs, and SpeechCOCO, we subsample up to 1,000 queries and limit the candidate pool to at most 10,000 instances. This strategy ensures computational tractability while maintaining a representative evaluation of model performance, consistent with common practices in large-scale retrieval benchmarks:

- **Audio Classification.** Datasets include ESC-50 (Piczak, 2015), UrbanSound8K (Salamon et al., 2014), NSynth (Engel et al., 2017), Speech Commands (Warden, 2018), and CREMA-D (Cao et al., 2014), evaluating the ability to recognize discrete acoustic events and sound categories.
- **Cross-modal Audio Retrieval.** This task evaluates alignment between audio and other modalities, including text-to-audio retrieval with Clotho (Drossos et al., 2020) and SoundDescs (Koepke et al., 2023), audio-video alignment with AVE (Tian et al., 2018), and audio-image retrieval with SpeechCOCO (Havard et al., 2017).
- **Audio Temporal Grounding.** Using the TUT Sound Events 2017 dataset (Mesaros et al., 2016), models must localize specific acoustic events within continuous audio streams.

Text Tasks. Recognizing that standard text retrieval tasks often fail to challenge modern information retrieval systems (Lu et al., 2026b), MMEB-V3 introduces more demanding scenarios involving instruction following, reasoning, and long-context understanding:

- **Instruction Following Retrieval.** FollowIR (Weller et al., 2025) and InfoSearch (Zhou et al., 2025) evaluate whether models can retrieve documents satisfying complex instructions and constraints.
- **Reasoning Retrieval.** BRIGHT (SU et al., 2025) and R2MD (Li et al., 2025) require logical inference and multi-hop reasoning beyond simple keyword matching.
- **Long-context Retrieval.** LongEmb (Zhu et al., 2024) evaluates retrieval from long documents and extended contexts.
- **Multi-condition Retrieval.** MultiConIR (Lu et al., 2025) measures the ability to satisfy multiple retrieval constraints simultaneously.

- **General Text Retrieval.** nanoBEIR (Thakur et al., 2021), a curated subset of BEIR, provides a compact yet diverse benchmark for semantic retrieval.

Agent Tasks. MMEB-V3 further evaluates agentic capabilities such as tool selection, GUI interaction and agent memory retrieval:

- **Tool Retrieval.** Tool-REX (Lu et al., 2026a) contains over 43,000 tools (e.g., APIs and code functions) with structured metadata, evaluating the ability to retrieve appropriate tools for user intents.
- **GUI Agent Trajectory Retrieval.** GAE-Bench (Zhang et al., 2025b) evaluates retrieval over multimodal GUI interaction trajectories represented by screenshots and structured actions.
- **Agent Memory Retrieval.** LMEB (Zhao et al., 2026) is a large-scale benchmark suite for long-horizon memory modeling. We select four representative tasks covering diverse memory types: *episodic memory* (KnowMeBench (Wu et al., 2026)), *dialogue memory* (REALTALK (Lee et al., 2025)), *semantic memory* (PeerQA (Baumgärtner et al., 2025)), and *procedural memory* (DeepPlanning (Zhang et al., 2026)). These tasks are used to evaluate whether embedding models can retrieve and utilize relevant memory under diverse semantic structures and temporal dependencies.

Omni-modality Semantic Equivalence Tuples Cross-modal retrieval is an information retrieval setting where a query in one modality is used to retrieve semantically aligned instances from a different modality. Despite recent progress in omni-embedding models that learn shared representation spaces across modalities, existing evaluations remain limited in measuring fine-grained, modality-aware alignment. To address this, MMEB-V3 introduces the concept of **Omni-modality Semantic Equivalence Tuples**, which is defined as a set of aligned instances $\{x^T, x^I, x^V, x^A\}$ with each instance encoding the same semantic meaning in one of the four different modalities. Under this setting, cross-modal retrieval is defined as the task of retrieving the correct element from a shared (flattened) pool of OmniSET that (1) is semantically equivalent to the input instance and (2) belongs to the specified target modality in the instruction.

- **OmniSET** contains 100 queries with human-verified hard negatives to increase discriminative difficulty. Each query is further expanded into 12 directional tasks, corresponding to the 12 ordered input-target modality pairs $(m_{in} \rightarrow m_{target})$ where $m_{in} \neq m_{target}$ and $m \in \{T, I, V, A\}$.

A.2.1 Construction of OmniSET

We introduce **OmniSET (Omni-modality Semantic Equivalence Tuples)**, a dataset designed to construct semantically equivalent instances across different modalities. The goal is to create tuples that share the same underlying semantics while differing in modality, enabling controlled evaluation of cross-modal alignment and transfer. Built upon the widely used MSCOCO dataset (Lin et al., 2015), which provides diverse real-world images with human-annotated captions, we extend the original bimodal (image–text) setting into a unified four-modality framework. Specifically, for each selected instance, we construct an *Omni-modality Semantic Equivalence Tuples (OmniSET)* consisting of aligned representations in text, image, video, and audio modalities.

Our construction consists of two main components. First, we curate a set of approximately 200 high-quality query samples, each paired with human-verified hard negatives to ensure fine-grained semantic discrimination. Second, we augment the original modalities by synthesizing additional ones: motion videos are generated from images using Google Veo-3.1, and speech audio is generated from captions using Gemini-2.5-Flash-TTS. This results in semantically aligned multi-modal tuples with consistent content across modalities. The overall construction pipeline is illustrated in Figure 5, and the prompt templates used for video and audio generation are provided in Figure 6.

The hard negative construction process is performed as follows:

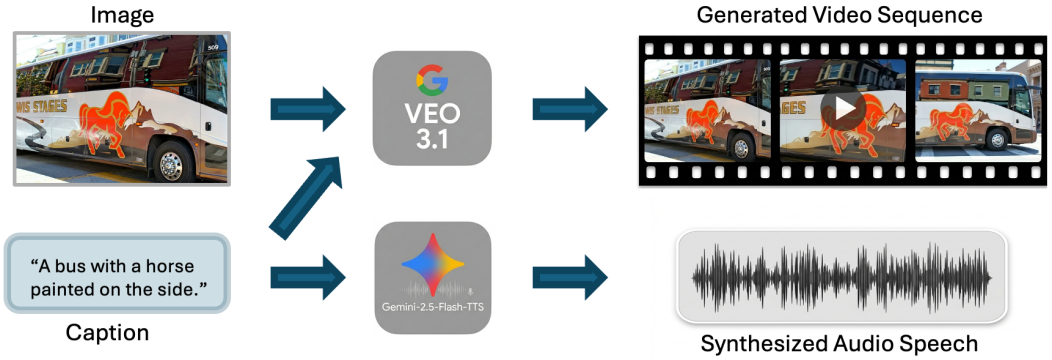


Figure 5: Construction pipeline of OmniSET.

1. We filter out images with restrictive licenses (e.g., Attribution-NoDerivs and Attribution-NonCommercial-NoDerivs) that prohibit derivative content.
2. For each remaining image, we extract object annotations and compute the number of unique object categories.
3. Given a reference image, we rank all other images by the number of shared object categories and select the top 30 as candidate hard negatives.
4. We sample approximately 100 images as query instances.
5. We manually inspect the candidates to remove near-duplicates or overly similar samples, resulting in 15–20 hard negatives per query.
6. Finally, for each positive instance and its associated negatives, we generate videos and audio to form complete OmniSET across all four modalities.

Video Generation Prompt

```
Generate a video based on the following image and description:
[Reference Image] + [caption]
```

Audio Generation Prompt

```
Read aloud the following sentence in a natural and expressive way,
in a warm and friendly tone:
[caption]
```

Figure 6: Prompt templates for video and audio generation.

We evaluate cross-modal retrieval by formulating queries and targets in different modalities. Each query is provided in one modality, and the model is required to retrieve a semantically matching instance from a specified target modality.

Given four modalities (text, image, video, and audio), each modality can query the other three, resulting in a total of 12 directed retrieval tasks (e.g., image $\rightarrow$ text, image $\rightarrow$ video, image $\rightarrow$ audio).

Shared candidate pool. For each query, all 12 retrieval directions share the same candidate pool, which contains the corresponding hard negatives across all four modalities. Therefore, regardless of the query direction, the model must select the correct target from a unified mixed-modality pool. This design ensures that performance differences across directions are not due to variations in candidate sets.

Avoiding trivial matching. A limitation of our dataset construction is that generated videos are highly consistent with their source images, which may simplify certain directions

such as I→V and V→I. To mitigate this, the input instance is explicitly included in the candidate pool as a distractor, while same-modality retrieval (e.g., I→I and V→V) is excluded. This prevents trivial matching based on instance identity.

A.2.2 Impact of Synthetic Data and Potential Modality Bias

A key design choice in OmniSET is the use of synthetic data to complete modality coverage: videos are generated from images, and audio is generated from text captions. While this enables controlled construction of semantically aligned tuples across modalities, it may introduce systematic biases that affect certain cross-modal directions.

In particular, the generation process creates asymmetric dependencies between modalities. Video samples are directly derived from images, and audio samples are derived from text. As a result, modality pairs such as image–video and text–audio may exhibit higher intrinsic similarity than other cross-modal pairs. This effect can partially explain the unusually strong performance observed in directions such as I→V and A→T, where retrieval may benefit from shared generation artifacts rather than purely learned cross-modal alignment.

More broadly, this construction may introduce a form of *modality preference*, where certain modality pairs are more tightly coupled due to data generation rather than model capability. This could potentially amplify observed behaviors such as modality bias or directional asymmetry. For example, when a query modality is closely aligned with a generated modality, retrieval may favor that modality even when it is not explicitly specified as the target.

However, several aspects of our design mitigate the extent to which these biases affect our conclusions. First, all retrieval directions share a unified mixed-modality candidate pool, ensuring that performance differences are not caused by variations in candidate sets. Second, trivial matching is explicitly avoided by including the source instance as a distractor and excluding same-modality retrieval. Third, the key phenomena observed in our analysis—such as the failure of instruction-constrained retrieval and the misalignment of instruction-induced shifts—are consistent across multiple models and modality directions, including those not directly affected by synthetic generation.

We therefore view these biases as an inherent trade-off of controlled multi-modal construction: while synthetic data may strengthen certain modality pairs, it does not fully account for the broader patterns observed in our experiments. Nevertheless, future work could further investigate this issue by incorporating human-annotated multi-modal data or by disentangling synthetic and real modality pairs through targeted ablation studies.

A.3 Detailed Scores

We provide complete per-dataset evaluation results for each task .

Image tasks. Results are reported in Table 5.

Video tasks. Results are reported in Table 6.

VisDoc tasks. Results are reported in Table 7.

Audio tasks. Results are reported in Table 8.

Text tasks. Results are reported in Table 9.

Agent tasks. Results are reported in Table 10.

A.4 Visualization and Analysis

In this section, we provide additional visualizations to complement the analysis in the main text. These figures illustrate both the global structure of the embedding space and the behavior of instruction-induced shifts across different models.

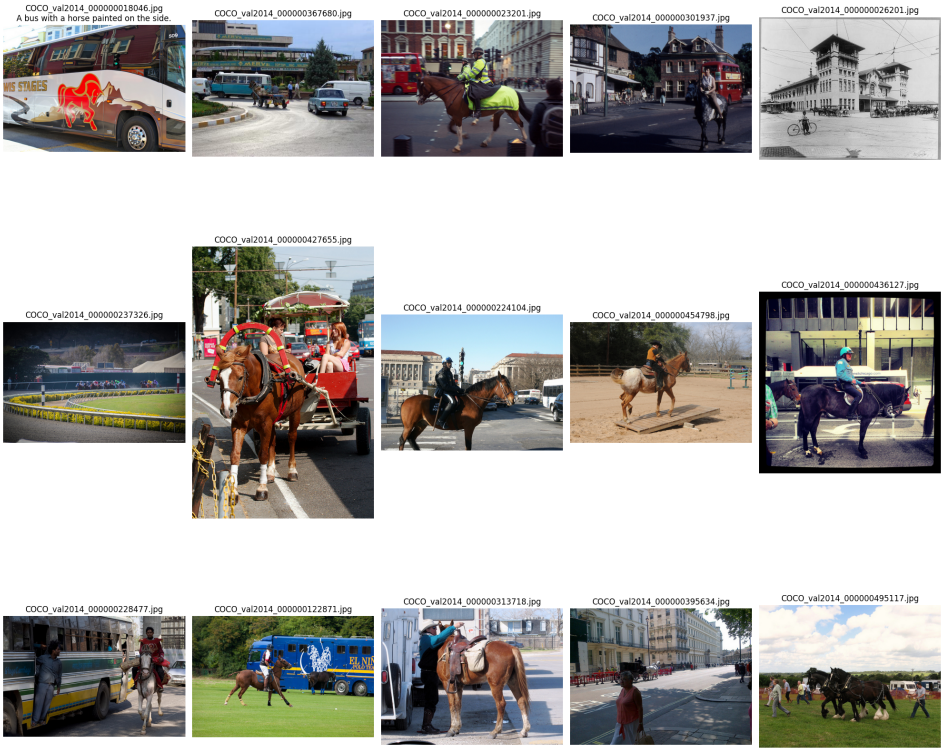


Figure 7: Example Hard Negative Set 18046 from OmniSET. **Query Image Caption**: A **bus** with a **horse printed** on the side. **The first image is the query input image*

Instruction-induced changes in query–target distance. Figure 10 presents the change in cosine distance between queries and target modalities after instruction augmentation for `WAVE` and `Qwen3-VL-Embedding-8B`. Compared to `omni-embed-nemotron-3b` in the main text, `WAVE` exhibits consistently small changes across all directions, indicating that instruction augmentation has limited influence on the query representation. In contrast, `Qwen3-VL-Embedding-8B` shows moderate variations, but the improvements are sparse and inconsistent. In most directions, the distance to the target modality either remains unchanged or increases, suggesting that instruction signals do not reliably improve target alignment. These observations are consistent with the findings in Section 5, where we show that both insufficient sensitivity and misaligned shifts contribute to retrieval failures.

Embedding space geometry across models. Figure 11 visualizes the embedding distributions of different modalities using t-SNE. For `omni-embed-nemotron-3b`, embeddings form relatively well-separated clusters corresponding to each modality, indicating a structured representation space. `WAVE` exhibits a more entangled distribution, particularly between audio and video modalities, reflecting its focus on audio–visual representation learning. In contrast, `Qwen3-VL-Embedding-8B` shows less clearly separated clusters, with significant overlap among modalities. This difference in geometry suggests that models vary in how modality information is encoded, which in turn influences their behavior under modality-constrained retrieval.

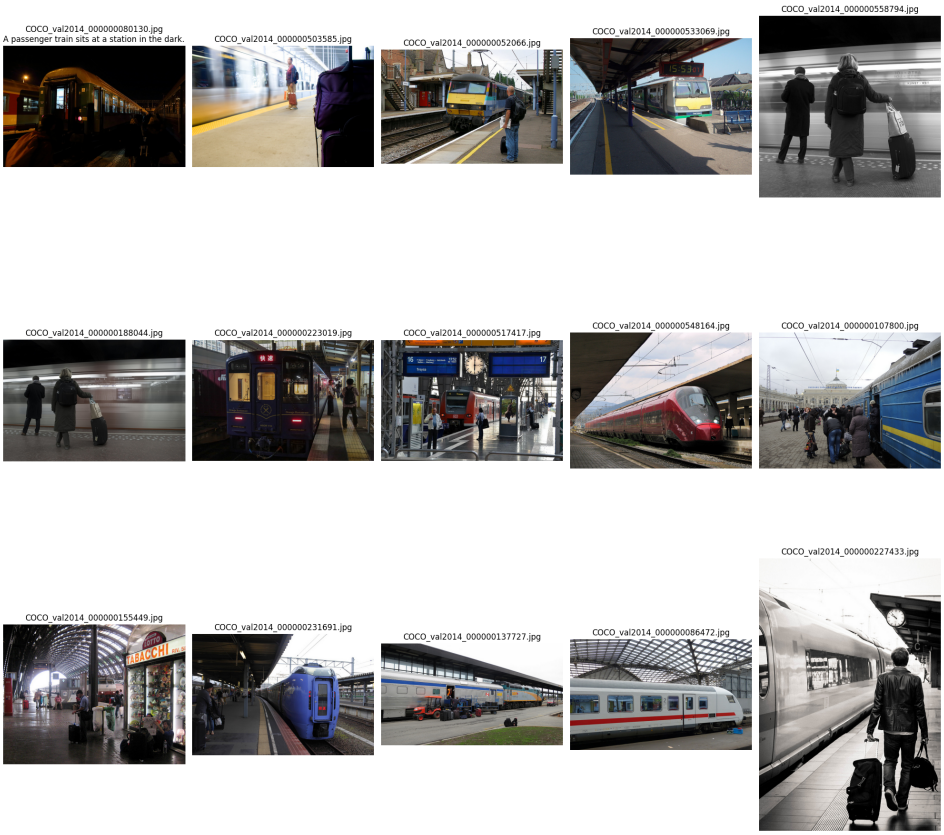
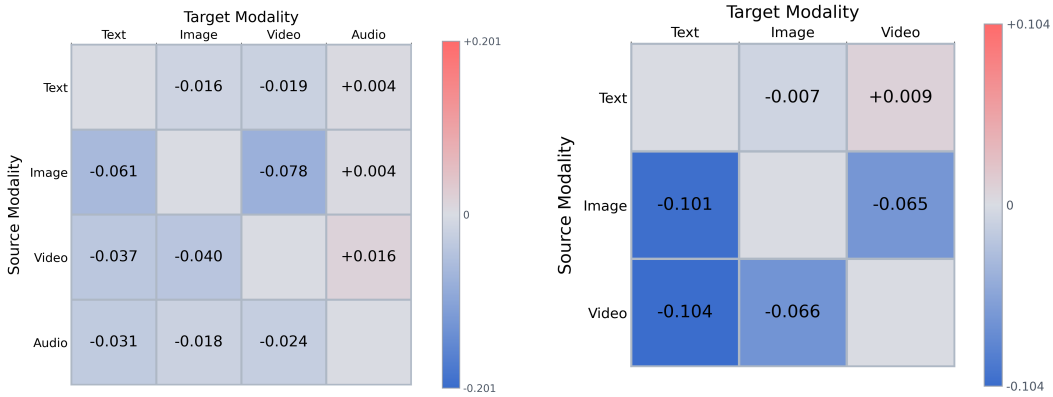


Figure 8: Example Hard Negative Set 80130 from OmniSET. **Query Image Caption**: A passenger train **sits** at a station **in the dark**. **The first image is the query input image*

Instruction-induced shifts across query modalities. Figure 12 further illustrates the effect of instruction augmentation on query embeddings for `omni-embed-nemotron-3b`, using video and audio queries as examples. Across different source modalities, instruction augmentation consistently perturbs the query representations. However, these shifts are not consistently directed toward the intended target modality. Instead, the updated query embeddings may move toward neighboring modality clusters or follow directions that are unrelated to the target. This observation provides additional qualitative evidence that, although the model is sensitive to instruction signals, the induced shifts do not reliably support target-oriented alignment.

Figure 9: Example Hard Negative Set 3067 from OmniSET. **Query Image Caption**: A **cat** sniffing a **sandwich** by an **apple** and a **phone**. **The first image is the query input image*



(a) WAVE. (b) Qwen3-VL-Embedding-8B.

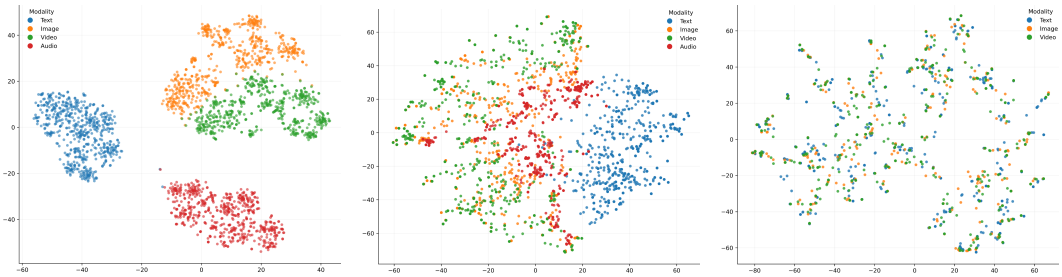
Figure 10: **Instruction-induced changes in query–target distance across models.** Each cell shows the change in cosine distance to the target modality after instruction augmentation, computed relative to the raw query. Negative values indicate that the instruction-augmented query moves closer to the target modality, while positive values indicate increased distance.

Dataset	Qwen3-VL -Embedding(2B)	Qwen3-VL -Embedding(8B)	VLM2Vec -Qwen2VL (7B)	VLM2Vec -V2.0 (2B)	GME (7B)	WAVE (7B)	Omni-Embed -Nemotron (3B)	LCO-Embedding -Omni (7B)	E5-Omni (3B)	E5-Omni (7B)
Avg - Image (37 tasks, Hit@1)	69.5	72.1	63.6	63.3	55.2	41.5	43.4	42.2	63.5	70.5
I-CLS (10)	63.6	65.2	62.8	62.9	57.6	50.1	48.3	57.3	60.7	68.2
I-QA (10)	71.5	76.8	56.5	56.4	34.6	26.3	19.9	15.9	62.2	70.3
I-RET (13)	68.4	69.0	69.4	69.6	71.2	42.1	56.1	48.8	65.0	69.7
I-VG (4)	82.9	87.7	81.9	77.1	59.5	56.1	48.9	48.8	68.6	79.6
ImageNet-1K	71.2	75.3	80.2	80.8	64.7	43.8	61.6	69.9	72.0	79.9
N24News	65.2	64.4	79.5	73.0	50.3	43.7	36.7	37.4	73.2	77.9
HatefulMemes	59.7	68.4	70.3	55.8	53.9	51.1	48.8	46.8	56.6	69.9
VOC2007	84.1	84.8	80.6	84.9	80.1	79.9	53.2	74.5	75.2	84.5
SUN397	67.9	64.7	77.3	70.9	69.4	64.1	55.1	70.2	71.0	76.0
Place365	37.8	37.6	37.2	36.1	39.1	34.0	31.9	39.0	41.0	44.5
ImageNet-A	59.9	61.6	57.6	47.6	40.6	25.2	42.0	50.7	42.3	60.5
ImageNet-R	91.1	92.0	74.2	89.3	83.9	69.1	82.8	91.0	85.2	86.6
ObjectNet	78.7	78.2	40.5	65.1	69.0	69.1	46.6	66.2	64.7	69.8
Country211	20.7	24.5	30.2	25.8	24.6	21.0	24.4	27.2	25.9	32.3
OK-VQA	65.9	74.7	56.8	51.7	33.1	32.7	17.7	14.0	62.7	70.6
A-OKVQA	58.8	66.6	47.5	44.0	20.8	24.8	12.5	9.6	52.3	59.7
DocVQA	93.9	95.2	88.8	90.1	41.1	19.7	17.1	16.9	86.6	93.6
InfographicsVQA	69.5	81.5	59.0	59.1	20.5	16.9	8.5	7.4	55.3	72.7
ChartQA	61.4	70.4	56.6	48.1	17.7	13.3	13.6	12.5	48.2	65.1
Visual7W	62.8	65.9	52.7	52.8	22.2	19.2	7.3	9.3	55.7	63.6
ScienceQA	71.4	75.6	38.2	38.1	28.2	21.6	24.8	22.6	47.9	56.4
VizWiz	57.2	59.5	39.6	43.3	38.9	33.1	34.2	31.1	51.8	56.1
GQA	85.9	88.6	54.4	65.4	76.8	50.6	28.0	11.6	78.5	78.9
TextVQA	88.1	90.2	71.6	71.6	46.4	30.6	35.3	23.9	82.8	86.3
VisDial	74.8	66.8	81.6	82.7	60.9	21.8	51.3	51.0	77.8	80.2
CIRR	55.5	56.7	51.4	57.3	54.9	31.9	13.2	8.1	36.8	49.7
VisualNews_t2i	68.4	75.7	80.3	74.7	79.5	45.7	60.2	63.7	69.1	74.1
VisualNews_i2t	74.0	81.9	81.2	78.3	83.6	40.2	58.5	59.2	76.9	82.1
MSCOCO_t2i	75.4	74.0	77.5	75.9	71.5	64.0	59.6	64.8	74.8	76.3
MSCOCO_i2t	71.2	73.7	73.6	71.1	57.4	55.8	55.8	55.6	71.1	74.9
NIGHTS	67.5	68.2	67.5	68.4	67.6	58.8	64.9	65.2	66.0	64.6
WebQA	88.7	89.3	88.3	90.6	91.3	49.7	90.2	52.7	90.1	88.6
FashionIQ	35.3	31.8	16.8	19.6	37.6	7.2	8.4	4.0	16.6	21.4
Wiki-SS-NQ	78.9	80.3	62.1	67.6	78.4	50.5	87.2	60.8	74.6	83.2
OVEN	70.2	70.7	66.6	64.8	75.6	57.3	72.3	64.4	75.2	78.9
EDIS	87.3	88.9	85.9	84.2	96.1	54.9	81.1	73.8	84.7	90.4
MCMR	42.0	38.0	0.9	4.1	27.3	8.9	26.1	11.0	31.9	41.1
MSCOCO	66.0	75.3	75.4	66.2	31.4	30.3	31.7	33.3	46.5	62.0
RefCOCO	89.0	93.8	87.1	87.0	61.2	57.4	55.8	55.1	76.4	84.7
RefCOCO-Matching	93.1	93.7	84.4	86.3	78.7	83.4	65.8	60.2	80.3	92.0
Visual7W-Pointing	83.5	87.8	80.7	69.0	66.5	53.4	42.3	46.6	71.1	79.5

Table 5: Detailed Scores on Image Tasks.

Dataset	Qwen3-VL -Embedding(2B)	Qwen3-VL -Embedding(8B)	VLM2Vec -Qwen2VL (7B)	VLM2Vec -V2.0 (2B)	GME (7B)	WAVE (7B)	Omni-Embed -Nemotron (3B)	LCO-Embedding -Omni (7B)	E5-Omni (3B)	E5-Omni (7B)
Avg - Video (18 tasks, Hit@1)	55.9	58.6	33.8	34.7	38.4	43.1	41.4	48.4	48.9	50.8
V-CLS (5)	62.1	65.0	39.0	39.2	37.3	50.7	48.3	58.1	46.9	53.9
V-QA (5)	61.7	67.6	30.1	34.7	50.3	45.9	47.8	62.4	60.4	65.9
V-RET (5)	46.9	48.9	29.1	28.4	28.3	34.7	38.7	42.2	42.7	40.5
V-MR (3)	51.2	49.2	39.2	37.5	37.5	39.5	23.5	19.3	43.4	37.9
K700	55.2	51.7	35.4	38.2	39.6	54.2	43.4	53.6	48.5	57.4
SmthSmthV2	66.0	73.3	32.0	43.0	30.5	49.8	46.2	60.6	45.7	51.7
HMDB51	68.8	75.5	41.9	40.2	48.0	54.2	52.1	52.4	42.6	56.4
UCF101	83.7	85.5	62.1	60.0	54.4	74.1	66.6	75.7	60.5	71.0
Breakfast	36.7	38.8	23.6	14.8	13.9	21.2	33.3	48.3	37.2	32.8
MVBench	56.2	64.8	28.6	33.6	46.3	41.4	43.0	62.7	56.9	64.0
Video-MME	51.3	55.6	28.0	30.8	39.2	32.3	38.0	51.9	47.4	52.7
NExTQA	65.4	72.4	20.3	20.9	53.5	36.8	45.2	69.4	67.5	74.9
EgoSchema	62.6	64.8	22.2	35.0	46.4	47.8	54.8	49.8	59.0	57.2
ActivityNetQA	72.9	80.3	51.4	53.0	66.0	71.3	58.2	78.0	71.5	80.6
DiDeMo	47.9	44.9	29.6	30.0	26.3	29.8	41.7	43.3	47.7	42.4
MSR-VTT	48.3	50.1	34.7	27.8	31.8	37.3	40.1	44.7	45.6	44.8
MSVD	70.4	72.8	46.7	47.3	49.6	60.9	60.3	64.8	63.1	64.3
VATEX	42.5	45.4	25.4	26.2	24.7	34.2	32.8	36.5	36.5	32.3
YouCook2	25.1	31.1	9.0	10.6	8.9	11.4	18.5	21.9	20.5	18.7
QVHighlight	74.0	71.3	57.9	49.7	59.4	54.5	20.9	15.1	49.3	52.9
Charades-STA	31.5	26.4	18.6	20.1	13.9	27.8	12.8	11.0	19.8	19.4
MomentSeeker	48.2	49.9	41.0	42.9	39.3	36.4	36.8	31.7	61.0	41.3

Table 6: Detailed Scores on Video Tasks.



(a) `omni-embed-nemotron-3b`. (b) `WAVE`. (c) `Qwen3-VL-Embedding-8B`.

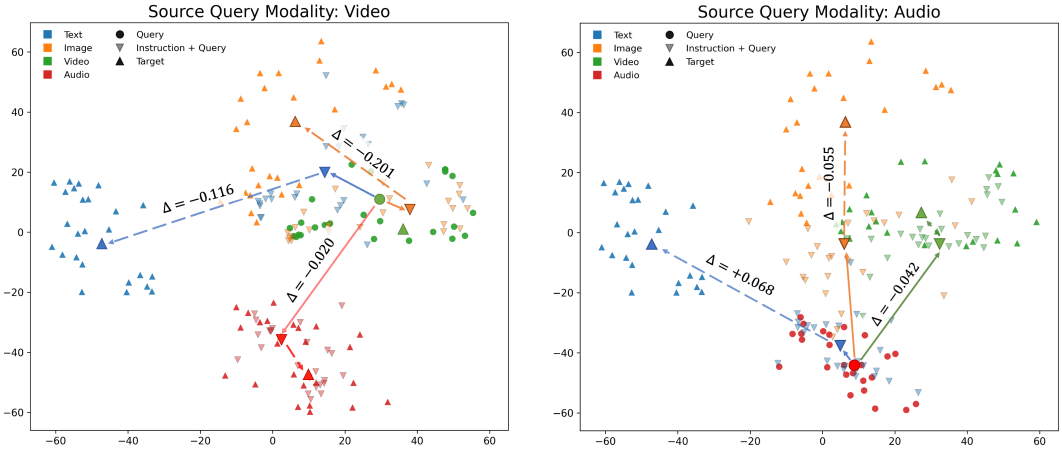
Figure 11: **Embedding space geometry across models.** t-SNE visualizations of embeddings from different modalities for three representative models.

Dataset	Qwen3-VL -Embedding(2B)	Qwen3-VL -Embedding(8B)	VLM2Vec -Qwen2VL (7B)	VLM2Vec -V2.0 (2B)	GME (7B)	WAVE (7B)	Omni-Embed -Nemotron (3B)	LCO-Embedding -Omni (7B)	E5-Omni (3B)	E5-Omni (7B)
Avg - VisDoc (24 tasks, NDCG@5)	70.6	70.9	32.6	68.6	75.2	42.8	71.1	68.7	73.5	75.4
ViDoRe-V1 (10)	81.2	82.1	20.0	74.4	89.6	52.3	84.8	80.1	84.5	87.3
ViDoRe-V2 (7)	58.5	54.6	9.2	44.6	55.5	30.7	51.6	54.1	57.3	58.3
VisRAG (6)	80.2	82.4	58.9	79.3	85.0	48.0	85.4	79.6	86.0	87.6
VisDoc-OOD (4)	41.9	41.8	48.1	62.0	44.4	23.3	35.2	38.6	43.1	44.2
ViDoRe_arxivqa	77.5	80.0	28.2	78.9	87.5	42.0	83.3	78.1	83.2	87.9
ViDoRe_docvqa	43.7	45.8	19.0	37.1	56.6	27.9	56.3	51.3	53.0	58.1
ViDoRe_infovqa	85.2	85.5	44.8	82.7	92.2	64.2	88.4	85.8	90.2	92.2
ViDoRe_tabfquad	94.9	94.8	17.0	87.8	92.7	68.0	92.3	92.1	93.4	93.0
ViDoRe_tatdqa	59.5	59.4	5.7	44.3	76.6	19.9	68.9	58.5	61.9	70.5
ViDoRe_shiftproject	79.4	81.0	1.6	61.0	95.6	43.1	79.7	71.4	83.2	84.7
ViDoRe_syntheticDocQA_artificial_intelligence	95.8	96.4	18.2	89.1	99.6	55.6	95.9	91.9	98.2	98.9
ViDoRe_syntheticDocQA_energy	88.2	89.2	23.9	86.3	95.7	67.2	92.7	89.4	92.5	94.1
ViDoRe_syntheticDocQA_government_reports	93.8	93.5	13.9	85.6	99.5	62.9	93.8	89.7	95.0	95.8
ViDoRe_syntheticDocQA_healthcare_industry	94.1	95.8	27.6	91.1	99.5	71.8	96.3	92.6	95.0	98.2
ViDoRe_esg_reports_human_labeled_v2	57.1	56.6	7.0	45.8	62.8	37.2	60.2	59.3	63.8	61.3
ViDoRe_biomedical_lectures_v2_multilingual	63.5	61.2	5.2	44.6	49.8	27.2	51.2	57.3	57.0	60.1
ViDoRe_economics_reports_v2_multilingual	57.7	50.4	13.8	42.3	53.9	32.0	44.2	55.2	56.5	53.5
ViDoRe_esg_reports_v2_multilingual	55.6	50.2	11.1	45.7	55.4	26.3	50.7	44.8	52.0	58.4
VisRAG_ArxivQA	79.6	78.7	52.9	76.7	87.7	31.6	83.2	77.0	84.8	88.1
VisRAG_ChartQA	79.2	86.4	69.0	84.2	81.3	56.7	89.2	85.2	87.6	87.8
VisRAG_MP-DocVQA	77.5	80.2	52.7	71.8	89.1	41.9	86.1	78.4	86.8	88.9
VisRAG_SlideVQA	91.8	92.3	72.8	91.4	94.7	68.0	95.2	90.7	95.8	96.0
VisRAG_InfoVQA	90.3	90.3	71.3	85.9	93.5	71.4	92.8	87.4	93.8	94.7
VisRAG_PlotQA	62.7	66.2	34.5	65.9	63.4	18.4	65.9	58.7	67.0	70.1
ViDoSeek-page	22.1	21.9	77.4	80.3	23.2	13.6	1.8	16.1	22.0	23.4
ViDoSeek-doc	84.5	84.6	54.2	80.3	83.9	48.0	83.1	80.8	83.1	83.9
MMLongBench-page	16.5	15.9	36.8	44.7	16.2	8.3	5.8	9.2	15.5	15.6
MMLongBench-doc	44.5	45.0	24.0	42.9	54.3	23.4	50.1	48.2	51.6	53.9

Table 7: Detailed Scores on VisDoc Tasks.

Dataset	WAVE (7B)	Omni-Embed -Nemotron (3B)	LCO-Embedding -Omni (7B)	E5-Omni (3B)	E5-Omni (7B)
Avg - Audio (11 tasks, Hit@1)	31.8	36.5	43.2	30.8	43.0
A-CLS (5)	52.3	53.7	64.3	41.4	63.0
A-RET (4)	12.9	18.8	19.9	22.0	23.5
A-TG (2)	21.5	29.1	36.8	21.6	32.2
NSynth	28.2	23.0	53.0	32.9	44.3
UrbanSound8K	50.2	54.4	56.0	48.9	57.4
ESC-50	83.3	76.5	66.1	40.0	68.8
SpeechCommands	43.6	87.0	84.8	50.4	86.9
CREMA-D	56.0	27.5	61.8	34.9	57.6
Clotho	14.8	12.3	14.0	17.5	19.3
SoundDescs	16.9	17.2	18.6	23.1	28.3
AVE	6.9	13.4	10.0	9.7	8.5
SpeechCOCO	7.4	32.2	37.0	37.7	37.9
TUTSound	38.4	53.7	66.9	39.5	59.3
TUTSound_hard	4.6	4.4	6.7	3.8	5.2

Table 8: Detailed Scores on Audio Tasks.



(a) Query modality: video.

(b) Query modality: audio.

Figure 12: Instruction-induced shifts in the embedding space for different source query modalities (`omni-embed-nemotron-3b`). Across different source modalities, instruction augmentation perturbs the query representations but does not consistently move them toward the target clusters.

Dataset	Qwen3-VL -Embedding(2B)	Qwen3-VL -Embedding(8B)	VLM2Vec -Qwen2VL (7B)	VLM2Vec -V2.0 (2B)	GME (7B)	WAVE (7B)	Omni-Embed -Nemotron (3B)	LCO-Embedding -Omni (7B)	E5-Omni (3B)	E5-Omni (7B)
Avg - Text (53 tasks, NDCG@5)	39.2	42.5	22.2	24.5	37.1	13.7	39.2	32.4	26.7	26.9
T-RR (20)	16.6	18.2	7.2	7.8	12.5	5.9	17.2	13.6	11.5	11.1
T-IF (9)	40.6	44.8	28.1	29.2	52.4	31.3	42.8	52.0	44.4	42.1
T-LR (6)	53.1	58.0	5.9	11.5	17.8	2.6	50.0	12.6	8.3	9.7
T-MR (5)	61.2	61.2	40.9	50.3	59.0	22.9	69.7	61.4	60.7	56.9
T-GR (13)	58.0	61.2	41.7	41.2	62.5	18.6	56.9	52.1	34.4	35.9
BRIGHT_aops	2.4	1.1	2.2	1.2	5.0	5.3	3.3	8.2	3.7	3.9
BRIGHT_biology	14.7	17.1	9.4	12.3	3.0	2.7	13.6	11.1	2.0	1.6
BRIGHT_earth_science	26.2	28.9	0.7	11.4	12.8	3.9	26.4	13.6	7.7	6.1
BRIGHT_economics	15.9	16.7	2.8	3.9	7.0	2.1	15.4	10.3	4.5	3.8
BRIGHT_leetcode	13.7	17.3	4.7	8.5	17.3	12.0	18.5	5.6	17.9	20.3
BRIGHT_pony	2.5	9.1	9.5	1.5	2.1	0.5	2.6	0.5	2.4	1.3
BRIGHT_psychology	18.6	12.0	3.0	4.4	11.0	2.8	18.1	8.5	5.6	4.2
BRIGHT_robotics	11.3	12.5	1.5	4.8	5.7	0.7	12.1	2.8	7.9	7.0
BRIGHT_stackoverflow	14.5	12.0	0.0	2.1	6.1	0.9	9.6	1.6	6.8	6.0
BRIGHT_sustainable_living	9.7	14.2	3.6	2.6	9.7	5.1	6.6	10.9	4.7	4.9
BRIGHT_theoremqa_questions	10.6	12.8	15.1	9.4	12.8	7.7	13.9	9.9	12.3	16.9
BRIGHT_theoremqa_theorems	12.8	21.2	3.2	4.7	17.0	1.9	23.0	8.6	11.2	11.6
FollowIR_core17-instructions	34.5	34.3	21.0	21.3	34.8	20.7	30.1	45.3	28.6	26.9
FollowIR_news21-instructions	21.8	20.4	9.7	4.6	13.3	11.1	31.6	26.7	25.3	24.9
FollowIR_robust04-instructions	39.5	44.0	21.1	24.7	30.6	11.0	33.5	40.2	32.1	27.4
InfoSearch_Audience-v1	31.3	34.5	25.0	17.2	61.8	31.9	27.4	48.6	29.1	34.2
InfoSearch_Clarity-v1	62.8	55.4	34.5	43.6	57.9	59.2	34.5	50.5	62.3	43.8
InfoSearch_Format-v1	38.2	46.2	10.2	20.4	47.5	7.2	52.5	47.3	49.9	49.2
InfoSearch_Language-v1	56.5	67.6	45.3	49.2	77.5	60.8	71.3	68.9	67.9	66.5
InfoSearch_Length-v1	44.2	55.9	40.0	43.9	75.4	38.0	56.9	73.3	56.4	58.6
InfoSearch_Source-v1	36.8	44.5	46.1	37.6	72.5	42.1	47.7	67.0	47.8	46.9
LongEmbed_2wikimqa	83.8	89.0	6.0	13.7	14.7	1.9	71.1	2.5	1.4	1.2
LongEmbed_narrativeqa	52.9	60.9	1.3	3.6	6.8	1.2	32.7	1.8	1.2	1.1
LongEmbed_needle	15.5	21.2	3.2	7.6	21.2	5.0	24.3	17.1	14.9	21.7
LongEmbed_passkey	32.8	38.2	12.7	11.3	26.7	1.1	35.9	31.8	27.7	31.4
LongEmbed_qmsum	37.5	42.3	4.7	11.9	9.8	2.4	40.2	8.4	1.5	1.5
LongEmbed_summ_screen_fd	95.9	96.3	7.5	20.9	27.4	3.8	96.0	13.8	3.1	1.3
MultiConIR_Books	62.1	60.9	35.0	47.1	56.8	27.7	64.0	56.1	52.1	52.5
MultiConIR_Legal Document	57.2	57.7	38.0	52.4	56.7	8.8	68.8	59.8	57.7	54.7
MultiConIR_Medical Case	60.0	61.4	35.8	43.9	52.8	17.5	65.7	59.4	57.7	51.6
MultiConIR_Movies	64.9	62.9	42.0	48.7	64.6	13.0	74.7	65.2	66.3	61.9
MultiConIR_People	61.8	62.9	53.4	59.2	64.1	47.3	75.1	66.2	69.8	64.0
NanoBEIR_NanoArguAna	49.4	45.0	32.4	36.1	54.3	21.6	42.8	25.7	53.3	49.9
NanoBEIR_NanoClimateFEVER	29.5	33.5	14.2	14.4	31.9	3.4	23.6	25.0	2.2	1.3
NanoBEIR_NanoDBPedia	60.6	61.1	51.7	48.4	60.5	23.6	53.8	65.8	38.2	36.2
NanoBEIR_NanoFEVER	89.2	89.8	55.9	52.3	87.1	10.5	75.4	80.0	18.6	19.0
NanoBEIR_NanoFiQA2018	43.7	53.2	20.1	21.0	58.8	4.8	54.4	43.2	15.4	24.0
NanoBEIR_NanoHotpotQA	73.8	77.4	53.6	50.0	84.5	15.3	92.0	62.2	73.0	56.9
NanoBEIR_NanoMSMARCO	53.6	64.9	28.1	38.6	65.5	19.0	53.4	45.7	47.0	44.4
NanoBEIR_NanoNFCorpus	39.5	39.9	31.8	29.5	39.9	11.7	31.8	34.8	10.8	13.2
NanoBEIR_NanoNQ	56.0	66.6	34.5	28.2	65.9	12.5	64.5	58.5	28.8	23.9
NanoBEIR_NanoQuoraRetrieval	94.9	96.0	93.4	92.8	95.2	87.1	91.5	90.5	92.9	87.5
NanoBEIR_NanoSCIDOCS	35.7	37.5	30.1	24.3	44.1	12.2	32.7	33.0	7.2	33.1
NanoBEIR_NanoSciFact	77.5	78.5	61.7	58.4	73.4	14.4	73.8	66.9	33.5	43.4
NanoBEIR_NanoTouche2020	50.1	52.0	34.2	41.6	52.0	5.6	50.3	46.3	25.7	33.6
R2MED_Bioinformatics	35.8	42.2	15.1	19.4	33.2	2.5	36.9	14.0	17.5	33.7
R2MED_Biology	14.7	17.2	9.3	12.3	3.2	2.7	13.9	11.0	2.0	1.6
R2MED_IIYi-Clinical	21.3	26.3	2.3	8.5	18.5	0.1	15.7	4.2	15.5	21.9
R2MED_MedQA-Diag	33.6	34.4	16.4	21.8	28.9	17.8	6.9	6.7	8.4	6.8
R2MED_MedXpertQA-Exam	11.5	17.4	6.0	5.6	5.3	0.5	5.3	6.5	5.3	6.9
R2MED_Medical-Sciences	10.1	13.1	3.5	4.7	9.3	0.8	27.3	33.1	25.8	20.0
R2MED_PMC-Clinical	21.4	37.1	13.5	14.0	32.2	1.0	21.7	8.7	28.8	25.7
R2MED_PMC-Treatment	30.4	37.4	21.4	28.3	42.0	1.1	14.2	13.9	27.6	34.9

Table 9: Detailed Scores on Text Tasks.

Dataset	Qwen3-VL -Embedding(2B)	Qwen3-VL -Embedding(8B)	VLM2Vec -Qwen2VL (7B)	VLM2Vec -V2.0 (2B)	GME (7B)	WAVE (7B)	Omni-Embed -Nemotron (3B)	LCO-Embedding -Omni (7B)	E5-Omni (3B)	E5-Omni (7B)
Avg - Agent (47 tasks, Hit@1)	39.3	38.4	19.7	28.7	35.6	11.3	36.5	27.8	36.9	36.7
Tool (35)	42.6	41.3	19.8	27.6	39.0	11.9	38.1	29.0	37.7	37.4
GUI (8)	30.4	33.5	21.4	36.2	30.0	11.8	32.0	25.0	35.6	38.0
Memory (4)	28.4	22.8	15.9	23.3	17.1	5.7	32.2	23.0	32.3	27.3
Tool-craft-math-algebra	84.6	83.9	57.1	65.0	84.6	17.5	85.0	63.9	75.7	81.4
Tool-craft-tabmwp	36.8	34.5	10.9	14.4	27.6	4.0	29.3	25.3	19.0	29.9
Tool-craft-vqa	52.0	54.5	23.0	43.5	60.5	12.5	65.5	34.5	47.0	50.5
Tool-gorilla-huggingface	26.4	28.4	18.6	20.2	24.6	6.4	26.6	20.2	25.8	19.2
Tool-gorilla-pytorch	11.6	9.3	0.0	4.7	14.0	2.3	11.6	16.3	20.9	14.0
Tool-gorilla-tensor	16.4	16.4	1.8	7.3	21.8	1.8	20.0	5.5	16.4	12.7
Tool-toolink	67.8	67.6	36.4	40.4	66.4	25.0	67.0	65.6	68.4	70.6
Tool-apibank	62.4	58.4	1.0	38.6	52.5	0.0	45.5	7.9	53.5	47.5
Tool-apigen	55.4	55.5	49.3	46.5	54.0	23.8	64.0	50.2	59.7	61.2
Tool-mnms	36.4	30.3	27.3	30.3	30.3	21.2	42.4	27.3	54.5	51.5
Tool-reversechain	61.0	57.5	22.5	51.5	54.5	0.5	50.0	18.5	64.0	56.5
Tool-rotbench	3.3	4.2	1.1	1.8	7.3	1.1	6.9	2.5	5.1	4.2
Tool-t-eval-dialog	24.0	20.0	26.0	22.0	34.0	0.0	14.0	20.0	16.0	24.0
Tool-t-eval-step	16.0	12.0	38.0	10.0	30.0	8.0	12.0	22.0	26.0	12.0
Tool-taskbench-daily	65.0	62.5	45.0	52.5	72.5	2.5	67.5	62.5	90.0	75.0
Tool-toolace	65.7	62.3	33.3	56.4	69.0	24.3	67.0	51.7	67.1	71.4
Tool-toolbench	40.6	36.0	6.0	24.8	17.4	4.6	13.3	16.9	29.2	28.9
Tool-toolemu	31.6	26.3	0.0	10.5	13.2	0.0	34.2	7.9	10.5	5.3
Tool-tooleyes	19.0	20.0	4.2	15.8	20.0	0.0	10.5	11.6	24.2	12.6
Tool-toollens	8.0	7.0	1.3	8.9	1.3	1.6	1.0	2.9	5.1	10.2
Tool-ultratool	50.6	41.2	0.4	4.6	40.4	5.2	43.6	41.2	44.0	43.8
Tool-autotools-food	50.0	50.0	9.1	4.6	18.2	0.0	13.6	4.5	9.1	18.2
Tool-autotools-music	12.5	37.5	0.0	0.0	0.0	0.0	3.1	0.0	0.0	0.0
Tool-autotools-weather	18.2	9.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.1
Tool-restgpt-spotify	60.0	67.5	27.5	12.5	20.0	0.0	25.0	0.0	20.0	30.0
Tool-restgpt-tmdb	37.0	38.9	0.0	0.0	11.1	0.0	16.7	9.3	9.3	13.0
Tool-appbench	71.9	84.4	65.6	65.6	81.2	71.9	87.5	59.4	87.5	84.4
Tool-gpt4tools	84.4	68.8	28.1	46.9	81.2	50.0	75.0	81.2	37.5	50.0
Tool-gta	14.3	21.4	14.3	0.0	35.7	0.0	42.9	28.6	28.6	14.3
Tool-taskbench-huggingface	60.9	39.1	4.3	56.5	56.5	8.7	52.2	52.2	47.8	56.5
Tool-taskbench-multimedia	70.0	62.5	32.5	65.0	87.5	42.5	77.5	60.0	92.5	82.5
Tool-metatool	52.5	51.5	32.0	44.0	54.0	26.5	49.0	48.5	46.0	51.0
Tool-tool-be-honest	51.1	49.1	47.1	46.0	55.4	23.1	54.3	35.4	54.0	52.0
Tool-toolalpaca	62.8	67.0	26.6	51.1	59.6	27.7	50.0	60.6	58.5	60.6
Tool-toolbench-sam	11.7	11.7	1.0	4.1	7.6	3.0	8.1	2.5	5.6	6.6
GAE-guiact-q2s	19.8	29.5	7.7	25.9	31.3	5.5	26.1	24.1	29.9	32.3
GAE-guiact-q2t	48.4	46.3	42.6	52.6	31.3	7.4	39.8	42.8	50.9	55.5
GAE-guiact-s2s	25.5	30.2	33.2	35.8	32.7	17.2	35.9	17.9	36.5	36.4
GAE-guiact-t2s	25.0	24.9	8.6	24.1	17.1	5.9	20.7	13.8	21.9	23.0
GAE-mind2web-q2s	27.1	37.7	12.2	43.0	42.1	10.9	41.7	33.0	43.0	44.9
GAE-mind2web-q2t	44.3	37.2	32.4	47.8	32.4	11.9	33.6	25.3	40.3	50.2
GAE-mind2web-s2s	30.2	35.2	32.4	37.1	37.1	28.0	37.1	25.9	38.3	37.7
GAE-mind2web-t2s	23.1	27.4	1.9	23.4	15.9	7.8	21.5	17.1	23.7	23.7
KnowMeBench	41.5	35.3	22.3	32.6	19.0	11.6	44.4	35.8	45.1	38.4
REALTALK	35.4	30.5	11.8	18.4	19.6	6.6	32.0	28.1	27.8	27.5
PeerQA	17.7	16.2	11.0	21.3	14.7	3.7	18.4	15.4	22.8	19.9
DeepPlanning	19.2	9.2	18.3	20.8	15.0	0.8	34.2	12.5	33.3	23.3

Table 10: Detailed Scores on Agent Tasks.