LOOK ACROSS LAYERS, LEARN THROUGH CHOICES: ENHANCING MULTIMODAL RETRIEVAL WITH UNIFIED REPRESENTATION AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models have recently made remarkable progress in embedding-based retrieval tasks. However, research on the joint optimization of multimodal representation and generation within a single architecture remains underexplored. Existing unified approaches exhibit a trade-off: they tend to substantially compromise representational capacity to preserve generative performance, or fail to establish a synergistic alignment between the two capabilities. To address these limitations, we propose **CLUE**, a **c**ross-**l**ayer **u**nified r**e**presentation and g**e**neration framework designed for holistic multimodal retrieval and comprehension. Specifically, CLUE introduces a Cross-Layer Embedding mechanism that integrates hierarchical semantic representations across Transformer layers, alleviating the information bottleneck of single-layer pooling. Furthermore, CLUE incorporates an Embedding Candidate Selection paradigm that leverages multimodal embeddings for candidate-level relevance discrimination through generative selection, establishing a more direct interaction between representation learning and generation. Extensive experiments on multimodal retrieval and comprehension benchmarks demonstrate that CLUE achieves state-of-the-art performance among the compared methods, substantially outperforming existing unified models on MMEB while effectively retaining robust generative capabilities.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have recently achieved remarkable progress in multimodal understanding, generation, and complex reasoning (OpenAI et al., 2024; Bai et al., 2025; Wang et al., 2025), providing strong general-purpose capabilities for downstream tasks such as visual question answering, content understanding, and multimodal reasoning. Meanwhile, MLLM-based universal multimodal representation learning has also advanced rapidly (Meng et al., 2025; Gu et al., 2026a; Sun et al., 2026; Liu et al., 2026b), demonstrating strong embedding capabilities for multimodal retrieval and semantic matching. However, most existing studies focus on optimizing multimodal representation as a separate capability, while effectively unifying high-quality representation and generation with direct interaction between the two remains underexplored.

Recent studies such as CAFe (Yu et al., 2025) and CREM (Liu et al., 2026a) have begun to explore unified representation and generation, taking initial steps toward integrating the two capabilities within a single MLLM. CAFe jointly optimizes contrastive and autoregressive objectives but provides limited interaction between the two capabilities. Recognizing this limitation, CREM further requires shared representation tokens to support text generation, offering only limited supervision for improving the discriminative representations required by retrieval. Moreover, existing unified approaches typically construct embeddings from the final Transformer layer, leaving complementary hierarchical semantics across different depths underexplored. To address these limitations, we propose **CLUE**, a cross-layer unified representation and generation framework for multimodal retrieval and comprehension. Inspired by the query-based information aggregation mechanism of Q-Former (Li et al., 2023), we introduce **Cross-Layer Embedding**, which employs cross-attention to extract and integrate hierarchical semantic representations across different Transformer layers. Unlike conventional approaches that construct embeddings solely from the final-layer hidden state, Cross-Layer Embedding adaptively models the contributions of representations at different levels,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

thereby exploiting complementary cross-layer semantics, alleviating the information bottleneck of single-layer pooling, and producing richer and more discriminative multimodal representations.

Furthermore, we introduce **Embedding Candidate Selection** to establish a more direct interaction between generation and representation learning. Specifically, multiple semantically related candidates are jointly presented within the generative context together with their corresponding embeddings, and generation is formulated as a multi-candidate selection task in which the model identifies the candidate that best matches the query. Rather than merely requiring representations to support text generation, Embedding Candidate Selection encourages the model to explicitly distinguish fine-grained semantic differences among competing candidates and leverages such choice-based supervision to enhance the embedding space. In this way, generation is not only preserved but also actively contributes to representation learning.

Extensive experiments on standard multimodal retrieval and comprehension benchmarks demonstrate the effectiveness of CLUE. On MMEB (Jiang et al., 2025), it achieves state-of-the-art retrieval performance among compared methods, substantially outperforming existing unified representation-generation models while maintaining robust multimodal comprehension and generative capabilities.

Our main contributions are summarized as follows:

- We introduce **Cross-Layer Embedding**, which integrates complementary semantic representations across Transformer layers via cross-attention to construct more informative and discriminative multimodal representations.
- We introduce **Embedding Candidate Selection**, which incorporates multimodal embeddings into generative multi-candidate selection to distinguish fine-grained semantic differences among candidates, providing stronger discriminative supervision for representation learning.
- We propose **CLUE**, a unified multimodal representation and generation framework. Experiments demonstrate that CLUE achieves state-of-the-art multimodal retrieval performance among the compared methods, while largely preserving the generative capability of the underlying MLLMs.

2 Related Work

2.1 Multimodal Representation Learning

Early vision-language models established strong foundations for multimodal representation learning through large-scale visual-textual alignment (Radford et al., 2021a; Li et al., 2022; Yu et al., 2022; Zhai et al., 2023). With the rapid development of MLLMs, recent studies further leverage their rich semantic understanding capabilities to construct universal multimodal embeddings for diverse retrieval and matching tasks (Jiang et al., 2024; 2025; Zhang et al., 2025). Building upon these advances, subsequent works improve multimodal representation learning from several perspectives, including scaling and refining training data (Zhou et al., 2025; Chen et al., 2025; Li et al., 2026), strengthening the learning signals of hard negatives (Lan et al., 2025; Xue et al., 2025), improving hard-negative mining strategies (Thirukovalluru et al., 2025; Gu et al., 2026b; Wu et al., 2026b), transferring discriminative knowledge from stronger models through distillation (Gu et al., 2025; Song et al., 2026), and incorporating explicit reasoning or Chain-of-Thought into embedding learning (Cui et al., 2026; Lan et al., 2026; Jiang et al., 2026). Despite these advances, most existing methods primarily focus on improving representation capability itself, leaving the generative capabilities of MLLMs and their interaction with representation learning underexplored.

2.2 Unified Multimodal Representation and Generation

Recent studies have explored unifying multimodal representation and generation within a single MLLM. CAFe (Yu et al., 2025) jointly optimizes contrastive and language modeling objectives to preserve both embedding and generative capabilities. CREM (Liu et al., 2026a) further introduces learnable compression tokens to support both retrieval and text generation. BToks (Sun et al., 2026) employs learnable bottleneck tokens and generative information condensation to improve semantic compression. ReMatch (Liu et al., 2026b) introduces generative matching over raw inputs and projected embeddings, but formulates matching as pointwise binary relevance prediction, where each query–candidate pair is independently judged as relevant or irrelevant. Although BToks and

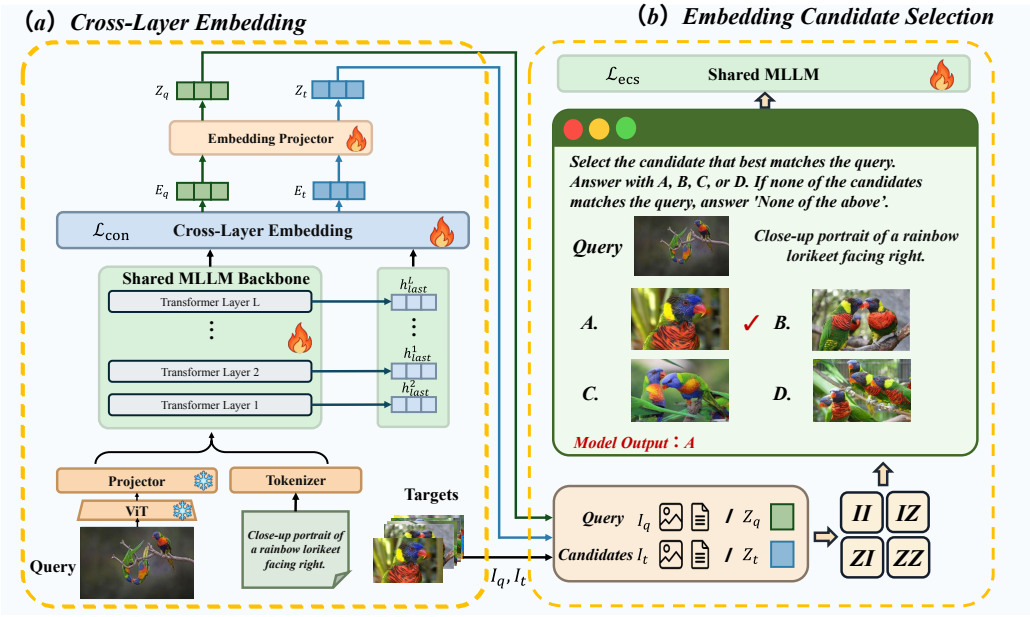


Figure 1: Overview of CLUE. (a) Cross-Layer Embedding aggregates the last-token hidden states from different Transformer layers to obtain multimodal representations E_q and E_t. (b) Embedding Candidate Selection incorporates the projected embedding vectors Z_q and Z_t into an MLLM-based candidate selection framework. I and Z denote the original multimodal inputs and projected embedding inputs, respectively, and different combinations indicate different pair types.

ReMatch strengthen the interaction between representation and generation, the interaction is still limited in how directly representation learning and generation supervise each other. Despite these advances, existing methods mainly connect representation and generation through joint objectives, shared tokens, or auxiliary generative supervision. In contrast, CLUE strengthens the integration of representation and generation by exploiting complementary cross-layer semantics and introducing multi-candidate generative discrimination for multimodal embeddings.

3 METHOD

We develop CLUE, a unified framework for multimodal representation learning and natural language generation. Figure 1 illustrates the overall framework of CLUE, which consists of two key components: Cross-Layer Embedding (CLE) and Embedding Candidate Selection (ECS). Specifically, CLE aggregates hidden representations across Transformer layers to construct multimodal embeddings, while ECS reformulates query-candidate matching as a generative multiple-choice task. In the following sections, we first introduce CLE in Sec. 3.1 and ECS in Sec. 3.2, then describe the autoregressive language modeling objective in Sec. 3.3, and finally present the overall training objective in Sec. 3.4.

3.1 CROSS-LAYER EMBEDDING

Existing MLLM-based multimodal representation learning methods typically adopt the hidden state of the last token from the final Transformer layer as the multimodal embedding (Jiang et al., 2025). This design provides a simple and effective interface for adapting autoregressive MLLMs to representation learning. Specifically, for an MLLM with L Transformer layers, the hidden state of the last token at the final layer, denoted as h^L_{last}, is commonly used as the final representation.

Despite its effectiveness, such a representation only relies on the final Transformer layer and does not exploit information distributed across different depths. Recent studies investigating the internal mechanisms of MLLMs have shown that multimodal information evolves throughout the Trans-

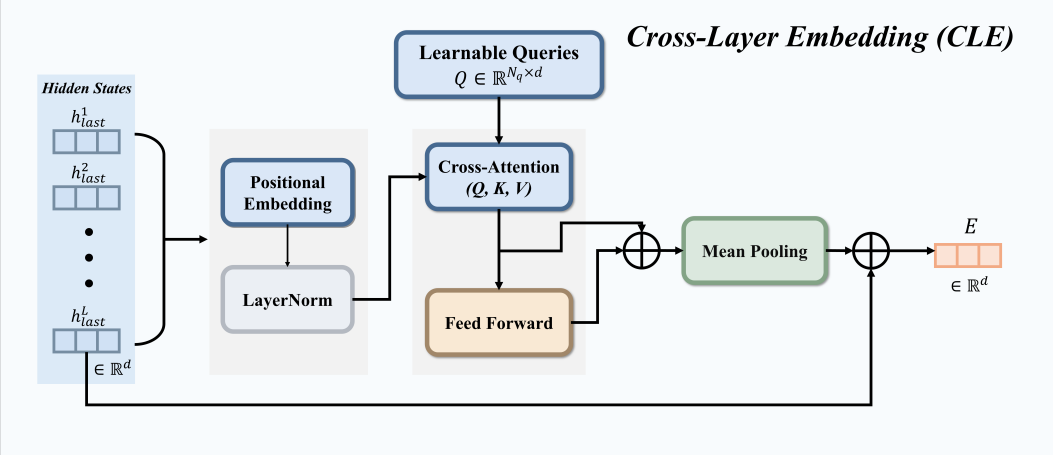


Figure 2: Overview of Cross-Layer Embedding (CLE). CLE aggregates the last-token hidden states from different Transformer layers through learnable queries and cross-attention, followed by feed-forward refinement and pooling to obtain the final embedding E. A residual connection from the original final-layer representation is used to preserve final-layer information.

former hierarchy, with different layers exhibiting distinct characteristics in vision-language interaction and semantic encoding (Tao et al., 2024; Zhang et al., 2025). These observations indicate that hidden states from different Transformer layers do not necessarily contain identical information, motivating us to exploit cross-layer information for more expressive multimodal representations.

Based on this motivation, we propose Cross-Layer Embedding, which extends conventional single-layer embeddings by aggregating information across Transformer layers. As illustrated in Figure 2, CLE introduces a cross-layer aggregation branch on top of the original final-layer representation. The branch consists of layer positional embeddings, learnable query-based cross-attention, a feed-forward network for further feature refinement, and residual fusion. Notably, rather than replacing the original final-layer representation, CLE preserves h^L_{last} while incorporating complementary information extracted from different Transformer depths.

Specifically, we first collect the last-token hidden states from all Transformer layers, denoted as $h^1_{\text{last}}, h^2_{\text{last}}, \ldots, h^L_{\text{last}}$, and organize them into a layer-wise representation sequence:

$$H = [h^1_{\text{last}}, h^2_{\text{last}}, \ldots, h^L_{\text{last}}]. \tag{1}$$

Before cross-attention, we inject layer positional embeddings into H and apply LayerNorm. Since hidden states from different Transformer layers may contain distinct multimodal information, we introduce N learnable query tokens to adaptively aggregate the resulting layer-wise representations across Transformer depths. The aggregation process is formulated as:

$$R = \text{CrossAttn}(Q, H, H). \tag{2}$$

The aggregated router tokens are further refined by a feed-forward network:

$$R_f = R + \text{FFN}(R) = [r_1, r_2, \ldots, r_N]. \tag{3}$$

We then apply mean pooling over the refined router tokens and combine the resulting cross-layer representation with the original final-layer embedding through residual fusion:

$$E = h^L_{\text{last}} + \frac{1}{N} \sum_{i=1}^{N} r_i. \tag{4}$$

This residual design preserves the original final-layer representation while incorporating information aggregated from the full Transformer hierarchy. Consequently, CLE extends the representation source of conventional MLLM embeddings from a single final layer to the entire Transformer depth, enabling the final representation to capture both final-layer and cross-layer information.

Following previous multimodal representation learning methods (Jiang et al., 2025), we adopt an instruction-based multimodal contrastive learning framework to optimize the embedding space. Given paired multimodal data, we construct query–target pairs and employ the standard InfoNCE loss with in-batch negatives:

$$\mathcal{L}_{\text{con}} = -\log \frac{\phi(E_q, E_{t^+})}{\phi(E_q, E_{t^+}) + \sum_{t^- \in \mathbb{N}} \phi(E_q, E_{t^-})}, \tag{5}$$

where $\mathbb{N}$ denotes the negative targets within the batch, and $\phi(\cdot, \cdot)$ denotes the matching score. We use cosine similarity: $\phi(E_q, E_t) = \exp(\cos(E_q, E_t)/\tau)$, where τ is the temperature hyperparameter. The representations E_q and E_t are generated by CLE.

To encourage different router tokens to capture complementary information from different Transformer layers, we introduce a lightweight diversity regularization on the refined router tokens R_f. Without explicit constraints, router tokens may collapse into redundant representations, limiting cross-layer aggregation. We therefore encourage complementary router representations by minimizing their squared pairwise cosine similarities:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N(N-1)} \sum_{i \neq j} \left(\frac{r_i^\top r_j}{\|r_i\|_2 \|r_j\|_2} \right)^2. \tag{6}$$

We further visualize the cross-attention weights averaged over queries across Transformer layers in Appendix C, providing insight into the layer-wise attention pattern.

3.2 Embedding Candidate Selection

Unified MLLMs for representation and generation are typically optimized with both contrastive and language modeling objectives. However, the two objectives remain largely separated: contrastive learning models query–target relations in the embedding space, while language modeling mainly operates on the original multimodal inputs. Consequently, learned embeddings are not explicitly involved in the generative discrimination process.

To better connect representation learning with the generative capability of MLLMs, we propose Embedding Candidate Selection (ECS). As illustrated in Figure 1, ECS reformulates representation-learning samples as a generative multiple-choice task. For each query, its paired positive sample and three hard negatives mined by mmE5 (Chen et al., 2025) form four randomly shuffled candidates, from which the MLLM autoregressively generates the best-matching option, i.e., A, B, C, or D. By comparing multiple challenging candidates within the same generative context, ECS provides explicit candidate-level discriminative supervision beyond embedding similarity.

To further involve learned representations, both the query and candidates can be represented by either their original multimodal inputs (I) or projected embeddings (Z), resulting in four input forms: II, IZ, ZI, and ZZ. Given a CLE representation $E \in \mathbb{R}^{d_e}$, we project it into the MLLM hidden space as:

$$Z = \mathcal{P}(E), \qquad Z \in \mathbb{R}^{d_{\text{model}}}, \tag{7}$$

where $\mathcal{P}$ is a lightweight projector shared by the query and candidates. Each Z is treated as a single soft token that replaces the corresponding original input within its query or candidate span, without additional position or modality embeddings. These input combinations expose projected embeddings to both original multimodal content and other projected embeddings, encouraging them to preserve semantics useful for generative discrimination.

A standard multiple-choice task always contains a correct option, which may encourage the model to simply select the relatively best candidate. We therefore introduce a None of the Above (NOTA) strategy: for a subset of samples, the positive candidate is replaced by an additional hard negative and the target answer is changed to `None of the above`. This requires the model to additionally determine whether a valid match exists. The resulting ECS objective is instantiated with the prompt template shown below.

Prompt Template for Embedding Candidate Selection:

USER: Select the candidate that best matches the query.
Answer with A, B, C, or D. If none of the candidates matches the query, answer "None of the above".

<QUERY>
{query input I_q}/{projected query embedding Z_q}
</QUERY>

<CANDIDATE A>
{candidate A input I_t^A}/{projected candidate A embedding Z_t^A}
</CANDIDATE A>
<CANDIDATE B>
{candidate B input I_t^B}/{projected candidate B embedding Z_t^B}
</CANDIDATE B>
<CANDIDATE C>
{candidate C input I_t^C}/{projected candidate C embedding Z_t^C}
</CANDIDATE C>
<CANDIDATE D>
{candidate D input I_t^D}/{projected candidate D embedding Z_t^D}
</CANDIDATE D>

ASSISTANT: A/B/C/D/None of the above

ECS is optimized directly with the native autoregressive prediction objective of the MLLM. Given the ECS input X_{ecs} and the target answer sequence $\mathbf{y}^* = \{y_1^*, \ldots, y_T^*\}$, the loss is defined as:

$$\mathcal{L}_{\text{ecs}} = -\sum_{t=1}^{T} \log p_\theta \left(y_t^* \mid X_{\text{ecs}}, y_{<t}^* \right). \tag{8}$$

Through this formulation, ECS couples projected embeddings with generative candidate selection, providing direct discriminative supervision for representation learning and strengthening the interaction between representation and generation.

3.3 Autoregressive Language Modeling

Adapting MLLMs for multimodal representation learning usually relies on contrastive objectives to optimize the embedding space. However, such representation-oriented optimization may interfere with the autoregressive generation behavior learned during pre-training and instruction tuning, leading to degradation of the model's original generative capability (Yu et al., 2025; Liu et al., 2026a). To preserve the generative ability during representation learning, we follow the standard autoregressive language modeling objective used in previous unified methods (Yu et al., 2025).

Specifically, given a multimodal instruction input X_{lm} and its corresponding target response sequence $\mathbf{y} = \{y_1, \ldots, y_T\}$, the autoregressive language modeling loss is formulated as:

$$\mathcal{L}_{\text{lm}} = -\sum_{t=1}^{T} \log p_\theta \left(y_t \mid X_{\text{lm}}, y_{<t} \right), \tag{9}$$

where X_{lm} denotes the multimodal instruction input, y_t denotes the target response token at position t, and $p_\theta(y_t \mid X_{\text{lm}}, y_{<t})$ represents the predicted probability of y_t conditioned on both the multimodal input and preceding response tokens. By retaining this autoregressive language modeling objective during training, the model can maintain its original generation capability while learning multimodal representations.

Nevertheless, we observe that representation-oriented training still affects the model's response behavior. Specifically, the generated responses tend to become shorter and less elaborative, which may reduce response diversity and limit the completeness of open-ended generation. The model also exhibits a stronger tendency to answer *Yes* in binary question scenarios. For selected VQA, classification, and captioning samples in the training corpus, we use Qwen3-VL-8B-Instruct to regenerate natural responses for the corresponding multimodal inputs and use them as language-modeling targets for joint training. Detailed data construction and processing procedures are provided in Appendix A.

3.4 Training Objective

Based on the aforementioned objectives, we optimize CLUE with

$$\mathcal{L} = \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{lm}}\mathcal{L}_{\text{lm}} + \lambda_{\text{ecs}}\mathcal{L}_{\text{ecs}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{10}$$

where $\mathcal{L}_{\text{con}}$, $\mathcal{L}_{\text{lm}}$, $\mathcal{L}_{\text{ecs}}$, and $\mathcal{L}_{\text{reg}}$ denote the loss functions for contrastive learning, autoregressive language modeling, Embedding Candidate Selection, and router diversity regularization, respectively. The corresponding λ coefficients control their relative contributions. Together, these objectives jointly optimize multimodal representation and generation, while ECS provides an additional connection between the two through generative candidate-level discrimination. The diversity regularization further encourages complementary information aggregation across router tokens.

4 Experiments

4.1 Experimental Setup

Training Datasets. We construct our training data from three sources. First, we adopt MMEB-train from VLM2Vec (Jiang et al., 2025) as the primary corpus for multimodal representation learning. For each positive sample, we incorporate hard negatives mined by mmE5 (Chen et al., 2025), which strengthen contrastive representation learning and provide challenging candidate options for our ECS objective. To preserve the generative and understanding capabilities of the underlying MLLM, we further incorporate Vision-Flan (Xu et al., 2024). For fair evaluation, we remove samples that overlap with MMEB and other evaluation benchmarks from Vision-Flan. In addition, for selected samples from these training subsets, we construct natural-response targets for autoregressive language modeling. Detailed data construction, filtering, benchmark decontamination, and natural-answer annotation procedures are provided in Appendix A.

Evaluation and Metrics. We evaluate multimodal retrieval performance on the MMEB benchmark (Jiang et al., 2025), which provides a comprehensive evaluation suite covering diverse multimodal embedding tasks. Following its standard split, the 36 evaluation datasets are divided into 20 in-distribution (IND) subsets and 16 out-of-distribution (OOD) subsets. For generative evaluation, we follow the evaluation protocol of VLMEvalKit (Duan et al., 2024) and conduct a comprehensive assessment on six established multimodal understanding benchmarks, including MMBench (Liu et al., 2024), MM-Vet (Yu et al., 2024), AI2D (Kembhavi et al., 2016), HallusionBench (Guan et al., 2024), MMMU (Yue et al., 2024), and MMStar (Chen et al., 2024). These benchmarks cover a broad range of multimodal capabilities, including visual perception, knowledge understanding, reasoning, and instruction following.

Implementation Details. We conduct experiments on MLLMs of different scales, including Qwen2-VL and Qwen3-VL, both of which employ unified multimodal architectures for processing textual and visual inputs. All models are fine-tuned with LoRA, with the rank $r = 32$ and scaling factor $\alpha = 64$. We train all models for 3,000 steps on 8 NVIDIA A100 GPUs with a global batch size of 1,024, following previous works for fair comparison. The learning rate is set to 1×10^{-4} with a cosine decay schedule, and the contrastive temperature is fixed at 0.02. For CLE, we employ 16 learnable queries with 16 attention heads for cross-layer aggregation. The four input forms II, IZ, ZI, and ZZ are uniformly sampled during training, and NOTA samples are constructed with a probability of 0.2. The loss coefficients are set to $\lambda_{\text{con}} = \lambda_{\text{lm}} = 1$ and $\lambda_{\text{ecs}} = \lambda_{\text{reg}} = 0.1$.

4.2 Multimodal Retrieval Capability

As shown in Table 1, CLUE demonstrates strong multimodal retrieval performance across different model scales. On the 2B scale, CLUE with Qwen3-VL achieves an overall score of 71.2, outperforming existing representation-oriented methods, as well as unified representation-generation baselines. When scaling to Qwen3-VL-8B, CLUE further improves the overall score to 74.1 and maintains competitive performance across different meta-tasks and both IND and OOD subsets. Notably, CLUE consistently outperforms existing unified methods such as CAFe (Yu et al., 2025) and CREM (Liu et al., 2026a) across model scales. We attribute these improvements to the two key

Table 1: Results on MMEB. "IND" denotes in-distribution datasets, while "OOD" denotes out-of-distribution datasets. Scores are averages of Precision@1 across datasets within each group. The best result is shown in bold. Rows highlighted in blue denote methods that explicitly jointly optimize multimodal representation and generation, which are the most comparable baselines to CLUE.

Model	Backbone	#Params	Per Meta-Task Score				Average Score		
			Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of Datasets →			10	10	12	4	20	16	36
CLIP (Radford et al., 2021b)	ViT-L	0.4B	55.2	19.7	53.2	62.2	47.6	42.8	45.4
∼2B Models									
VLM2Vec (Jiang et al., 2025)	Qwen2-VL	2B	59.0	49.4	65.4	73.4	66.0	52.6	60.1
VLM2Vec-V2 (Meng et al., 2025)	Qwen2-VL	2B	62.9	56.3	69.5	77.3	68.7	60.1	64.9
LLaVE (Lan et al., 2025)	Aquila-VL	2B	62.1	60.2	65.2	84.9	69.4	59.8	65.2
BToks (Sun et al., 2026)	Qwen2-VL	2B	64.3	59.8	68.8	77.4	–	–	66.0
Embed-RL (Jiang et al., 2026)	Qwen3-VL	2B	62.3	67.5	68.7	89.4	–	–	68.9
LaME (Wu et al., 2026a)	Qwen2-VL	2B	67.6	66.2	70.5	81.2	73.1	64.4	69.3
RGE (Liu et al., 2025)	Qwen2.5-VL	3B	64.4	67.8	70.2	**90.1**	**75.3**	63.7	70.1
ReMatch (Liu et al., 2026b)	Qwen2-VL	2B	65.8	65.9	70.1	83.3	72.8	64.7	69.2
CAFe (Yu et al., 2025)	LLaVA-OV	1B	59.1	49.1	61.0	83.0	64.3	53.7	59.6
CREM (Liu et al., 2026a)	Qwen2-VL	2B	65.8	60.7	68.3	78.9	70.8	61.5	66.7
CLUE (Ours)	Qwen2-VL	2B	66.9	66.5	70.9	81.7	72.0	66.9	69.8
CLUE (Ours)	Qwen3-VL	2B	**67.7**	**68.4**	**71.2**	86.5	73.6	**68.1**	**71.2**
∼8B Models									
VLM2Vec (Jiang et al., 2025)	Qwen2-VL	7B	62.6	57.8	69.9	81.7	72.2	57.8	65.8
mmE5 (Chen et al., 2025)	Llama-3.2-Vision	11B	67.6	62.8	70.9	89.7	72.3	66.7	69.8
LaME (Wu et al., 2026a)	Qwen2-VL	7B	70.0	71.0	73.1	85.8	–	–	73.0
TWN (Zhang et al., 2026)	Qwen3-VL	8B	**70.2**	**74.3**	70.8	87.1	–	–	73.4
ReMatch (Liu et al., 2026b)	Qwen2-VL	7B	68.2	71.8	**73.9**	88.4	**77.4**	68.3	73.3
CAFe (Yu et al., 2025)	LLaVA-OV	7B	65.2	65.6	70.0	**91.2**	75.8	62.4	69.8
CREM (Liu et al., 2026a)	Qwen2-VL	7B	68.3	69.4	72.9	86.1	75.6	67.8	72.1
CLUE (Ours)	Qwen3-VL	8B	**70.2**	72.6	73.4	89.7	76.9	**70.6**	**74.1**

components of CLUE: CLE, which enriches multimodal representations by aggregating complementary information across Transformer layers, and ECS, which further strengthens discriminative representation learning through generative candidate-level supervision.

4.3 Multimodal Generation Capability

Beyond representation quality, a unified multimodal model should preserve the general-purpose multimodal understanding and generation capabilities of the underlying MLLM. We therefore evaluate CLUE on six widely used multimodal comprehension benchmarks and compare it with the corresponding original backbones. As shown in Table 2, CLUE largely retains the generative capability across different model architectures and scales while substantially improving multimodal representation performance. Specifically, CLUE based on Qwen2-VL-2B achieves an average score of 53.6, closely matching the 53.8 of the original backbone. On Qwen3-VL, CLUE obtains 61.2 versus 62.7 at the 2B scale, and 72.3 versus 73.7 at the 8B scale, showing similarly limited degradation as the model scales up. We speculate that the slightly larger degradation on Qwen3-VL may be due to its stronger generative capability, making it more sensitive to interference from joint optimization.

These results demonstrate that CLUE can substantially enhance multimodal representations while largely preserving the original multimodal comprehension and generation capabilities across different backbone architectures and model scales. By jointly optimizing representation learning and autoregressive language modeling within the same MLLM, CLUE maintains a favorable balance between embedding quality and generation performance.

4.4 Ablation Study

We conduct ablation studies on Qwen3-VL-2B to evaluate the contribution of each component in CLUE. As shown in Table 3, retrieval-only training achieves an MMEB score of 70.7, while adding language modeling improves generation from 52.6 to 61.3 but reduces MMEB to 69.9, indicating interference between representation and generation objectives. Introducing CLE improves MMEB to 71.1 under retrieval-only training, and further incorporating ECS achieves 71.2 while maintaining strong generation performance. These results show that CLE and ECS together recover the retrieval degradation caused by joint training and further improve representation quality.

Table 2: Results on multimodal comprehension benchmarks. "AVG" denotes the average score across all benchmarks.

Model	MMB	MMVet	AI2D	Hallusion	MMMU	MMStar	AVG
			2B Models				
Qwen2-VL-2B	71.7	50.0	74.0	39.5	40.7	46.6	53.8
CLUE	71.5	51.9	71.4	39.1	40.7	46.7	53.6
Qwen3-VL-2B	75.9	62.2	76.8	53.7	50.2	57.4	62.7
CLUE	74.6	60.2	76.1	51.3	48.9	56.1	61.2
			8B Models				
Qwen3-VL-8B	85.1	74.7	85.6	58.8	66.7	71.5	73.7
CLUE	84.2	72.7	84.0	58.3	64.0	70.3	72.3

Table 3: Ablation study of the main components of CLUE. "Ret." and "Gen." denote retrieval and generation training, respectively.

Ret.	Gen.	CLE	ECS	Generation	MMEB
✓				52.6	70.7
✓		✓		52.7	71.1
✓	✓			61.3	69.9
✓	✓	✓		61.2	70.4
✓	✓	✓	✓	61.2	**71.2**

Table 4: Ablation study of Embedding Candidate Selection.

Setting	Generation	MMEB
ReMatch-style	61.0	70.8
ECS-style	61.2	**71.2**
Input: ZZ	61.1	70.7
Input: II/IZ/ZI/ZZ	61.2	**71.2**
Candidates: 8	61.2	71.0
Candidates: 4	61.2	**71.2**
w/o NOTA	61.1	70.6
w/ NOTA	61.2	**71.2**

We further compare ECS with a ReMatch-style generative matching objective (Liu et al., 2026b) under otherwise identical settings. As shown in Table 4, replacing ECS with the ReMatch-style objective reduces the MMEB score from 71.2 to 70.8. ReMatch relies on pointwise binary relevance judgments, where candidates are evaluated independently, whereas ECS jointly compares a positive candidate with multiple hard negatives in the same generative context, providing stronger comparative supervision for representation learning.

We also analyze several design choices of ECS in Table 4. Using all four input forms improves MMEB from 70.7 to 71.2 compared with ZZ alone, suggesting that combining original inputs with projected embeddings enables richer cross-form interaction and preserves useful semantics. Increasing the number of candidates from 4 to 8 has limited impact on MMEB performance. We speculate that additional hard negatives increase selection difficulty without providing substantially stronger supervision. Incorporating NOTA improves MMEB from 70.6 to 71.2, as it prevents the model from assuming that a correct candidate must always exist and encourages more reliable relevance judgment. Sensitivity analyses of key hyperparameters are provided in Appendix B.

5 Conclusion

We present CLUE, a unified framework for enhancing multimodal representation learning while preserving the generative capability of MLLMs. CLUE leverages cross-layer semantic information to enrich multimodal embeddings beyond conventional final-layer representations, while further coupling representation learning with generative discrimination through candidate-level selection. Together, these designs strengthen multimodal retrieval while largely preserving the general-purpose understanding and generation capabilities of the underlying MLLM. Experiments demonstrate that CLUE achieves strong performance on multimodal retrieval benchmarks while maintaining competitive multimodal comprehension and generative capabilities. We hope CLUE provides a step toward deeper integration of representation learning and generation in unified multimodal models.

AI USE STATEMENT

In this work, we used generative AI tools as a general-purpose writing assistant during the preparation of the manuscript. Specifically, these tools were used for proofreading, correcting grammatical errors, improving sentence clarity, and assisting with language refinement.

We did not use generative AI tools for developing research ideas, designing the methodology, conducting experiments, analyzing experimental results, or drawing scientific conclusions. All AI-assisted suggestions were carefully reviewed and revised by the authors. The authors verified the correctness and consistency of the final manuscript, including all technical descriptions, claims, and experimental results, and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on improving unified multimodal representation and generation and does not involve human-subject experiments or the collection of personally identifiable information. All datasets and benchmarks used in this study are publicly available research resources, and we use them in accordance with their intended research purposes and applicable licenses or terms of use.

As with other multimodal representation and generation models, our method may inherit biases, inaccuracies, or potentially harmful content present in the underlying pretrained models and training data. Improved representation and retrieval capabilities may also be applied in unintended settings, such as large-scale content retrieval or matching. Our work does not introduce mechanisms specifically designed for surveillance, user profiling, or other privacy-sensitive applications, and we encourage future deployments to consider appropriate safeguards, access controls, and dataset-specific ethical requirements.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide comprehensive details of our proposed framework, training data, and experimental settings in both the main text and the appendix.

Methodology. A detailed description of our method is provided in Section 3, including Cross-Layer Embedding (Section 3.1), Embedding Candidate Selection (Section 3.2), autoregressive language modeling (Section 3.3), and the overall training objective (Section 3.4). We additionally provide framework illustrations in Figure 1 and Figure 2 to clarify the overall architecture and the design of Cross-Layer Embedding.

Datasets. We employ publicly available datasets for both training and evaluation. The training data and evaluation benchmarks are described in Section 4.1, including MMEB-based representation learning data, mmE5-mined hard negatives, Vision-Flan-based generative supervision, and the benchmarks used for multimodal retrieval and comprehension evaluation. Detailed data construction, preprocessing, filtering, and benchmark decontamination procedures are further provided in Appendix A.

Experiment Details. Detailed implementation and evaluation settings are provided in Section 4.1. These include model backbones, LoRA configurations, training steps, batch size, learning rate schedule, contrastive temperature, Cross-Layer Embedding configurations, Embedding Candidate Selection settings, and loss coefficients. We also report the evaluation protocols for multimodal retrieval and generation benchmarks to facilitate reproducible comparison.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun

Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mmE5: Improving multimodal multilingual embeddings via high-quality synthetic data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8254–8275, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.433. URL https://aclanthology.org/2025.findings-acl.433/.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. URL https://arxiv.org/abs/2403.20330.

Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Mishra, Taipeng Tian, Qi Guo, et al. Think then embed: Generative context improves multimodal embedding. In *International Conference on Learning Representations*, volume 2026, pp. 2690–2709, 2026.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Geonmo Gu, Byeongho Heo, Jaemyung Yu, Jaehui Hwang, Taekyung Kim, Sangmin Lee, HeeJae Jun, Yoohoon Kang, Sangdoo Yun, and Dongyoon Han. Muco: Multi-turn contrastive learning for multimodal embedding model, 2026a. URL https://arxiv.org/abs/2602.06393.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. *arXiv preprint arXiv:2504.17432*, 2025.

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21378–21386, 2026b.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, June 2024.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-rl: Decoupled reinforcement learning for reasoning-driven multimodal embeddings. *arXiv preprint arXiv:2602.13823*, 2026.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models, 2024. URL https://arxiv.org/abs/2407.12580.

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.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. URL https://arxiv.org/abs/1603.07396.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. LLaVE: Large language and vision embedding models with hardness-weighted contrastive learning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13721–13735, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.740. URL https://aclanthology.org/2025.findings-emnlp.740/.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Ume-r1: Exploring reasoning-driven generative multimodal embeddings, 2026. URL https://arxiv.org/abs/2511.00405.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, 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, 2026. URL https://arxiv.org/abs/2601.04720.

Chunxu Liu, Jiyuan Yang, Ruopeng Gao, Yuhan Zhu, Feng Zhu, Rui Zhao, and Limin Wang. Reasoning guided embeddings: Leveraging mllm reasoning for improved multimodal retrieval, 2025. URL https://arxiv.org/abs/2511.16150.

Lihao Liu, Biao Yang, Yan Wang, Da Li, Jiangxia Cao, Yuxiao Luo, Xiang Chen, Xiangyu Wu, Wei Yuan, Fan Yang, Guiguang Ding, Tingting Gao, and Guorui Zhou. Crem: Compression-driven representation enhancement for multimodal retrieval and comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 5714–5724, June 2026a.

Qianying Liu, Xiao Liang, Zhiqiang Zhang, Yibo Chen, Xu Tang, Zhongfei Qing, Fengfan Zhou, Yao Hu, and Paul Henderson. Rematch: Boosting representation through matching for multimodal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16823–16833, June 2026b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. URL https://arxiv.org/abs/2307.06281.

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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey

Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim,

Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

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. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/radford21a.html.

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, 2021b. URL https://arxiv.org/abs/2103.00020.

Tingyu Song, Mingxin Li, Yanzhao Zhang, Dingkun Long, Chu Liu, Pengjun Xie, Yilun Zhao, and Shu Wu. Core: Improving compositional reasoning in mllm embedding via reranker distillation, 2026. URL https://arxiv.org/abs/2609.04083.

Siyu Sun, Jing Ren, Zhaohe Liao, Dongxiao Mao, Xiangyuan Ren, Yiyi Zhang, Haohua Zhao, Weixiong Lin, Jiang Shaohua, Liqing Zhang, and Yuchao Zheng. Bottleneck tokens for unified multimodal retrieval, 2026. URL https://arxiv.org/abs/2604.11095.

Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. Probing multimodal large language models for global and local semantic representations. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 13050–13056, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1142/.

Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhu Chen, and Bhuwan Dhingra. Breaking the batch barrier (b3) of contrastive learning via smart batch mining, 2025. URL https://arxiv.org/abs/2505.11293.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Peixi Wu, Biao Yang, Feipeng Ma, Bosong Chai, Bo Lin, Wei Yuan, Fan Yang, Tingting Gao, Hebei Li, and Xiaoyan Sun. Lame: Learning to think in latent space for multimodal embedding via information bottleneck, 2026a. URL https://arxiv.org/abs/2606.13061.

Yebo Wu, Feng Liu, Ziwei Xie, Zhiyuan Liu, Changwang Zhang, Jun Wang, and Li Li. Tsembed: Unlocking task scaling in universal multimodal embeddings, 2026b. URL https://arxiv.org/abs/2603.04772.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning, 2024. URL https://arxiv.org/abs/2402.11690.

Youze Xue, Dian Li, and Gang Liu. Improve multi-modal embedding learning via explicit hard negative gradient amplifying, 2025. URL https://arxiv.org/abs/2506.02020.

Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning, 2025. URL https://arxiv.org/abs/2503.19900.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, Aug 2022, 2022. URL https://arxiv.org/abs/2205.01917.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yu24o.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Longxiang Zhang, Weilong Dai, Guanghao Zhang, Hao Jiang, and Pipei Huang. Think when needed: Adaptive reasoning-driven multimodal embeddings with a dual-lora architecture, 2026. URL https://arxiv.org/abs/2605.14448.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9274–9285, June 2025.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. MegaPairs: Massive data synthesis for universal multimodal retrieval. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.935. URL https://aclanthology.org/2025.acl-long.935/.

A Dataset Construction, Cleaning, and Processing

Our training data is constructed to jointly support multimodal representation learning and generation. For representation learning, we use the MMEB training corpus augmented with hard negatives mined offline by mmE5 (Chen et al., 2025), together with a carefully cleaned version of Vision-Flan. For language modeling, we further construct natural-answer annotations using a stronger MLLM teacher. In this section, we describe the composition, cleaning and benchmark decontamination of the training data, as well as the construction of natural-answer supervision.

A.1 Representation Learning Data

Our representation-learning data is primarily built upon the MMEB hard-negative training corpus released with mmE5. Compared with the original MMEB training data, this version augments each query–positive pair with hard negatives mined offline by mmE5. Each sample typically contains multiple hard-negative candidates. During training, we randomly sample one mined hard negative for each query and combine it with the remaining in-batch negatives in the contrastive objective. Compared with relying solely on in-batch negatives, the hard examples mined by mmE5 provide stronger discriminative supervision for representation learning.

We additionally incorporate Vision-Flan after converting it into the same query–target format. Since Vision-Flan does not contain mmE5-mined explicit hard negatives, samples from this subset use only in-batch negatives. This also reduces the risk of introducing false negatives for its many short-answer tasks.

The final representation-learning mixture consists of 21 subsets spanning image classification, visual question answering, image–text retrieval, composed retrieval, and visual grounding. All subsets are assigned equal sampling weights, while dataset-specific sampling caps are applied to prevent large-scale datasets from dominating the training distribution. Table 5 summarizes the final training mixture. After preprocessing and deduplication, the complete representation-learning pool contains approximately 1.24M training instances.

A.2 Vision-Flan Processing and Benchmark Decontamination

The original Vision-Flan collection contains 186,103 samples from 191 visual instruction-following tasks. We first convert its conversation-style annotations into the query–target format used by our representation-learning pipeline, where the human instruction is treated as the query and the corresponding response as the positive textual target.

Vision-Flan contains both open-ended tasks with diverse responses and short-answer or binary tasks with highly concentrated answer distributions. Directly retaining all such samples may introduce substantial false negatives under contrastive learning, since different samples can share identical or highly similar correct answers. We therefore apply task-aware subsampling, retaining a larger sampling budget for long-form tasks while more aggressively subsampling short-answer and binary tasks.

We additionally remove duplicated samples and responses with abnormally long target sequences. In particular, samples are deduplicated according to their task identity, image, and normalized response. After format conversion, task-aware sampling, and basic deduplication, the number of Vision-Flan samples is reduced from 186,103 to 184,031.

Dataset-level Benchmark Decontamination. Vision-Flan aggregates a large collection of existing vision-language datasets, some of which directly overlap with the source datasets used by our evaluation benchmarks. To prevent training–evaluation leakage, we first perform *dataset-level decontamination*. If a benchmark source can be clearly identified as a complete Vision-Flan task and directly overlaps with our evaluation data, the corresponding task is entirely excluded before sample-level matching.

In particular, all MMEB OOD benchmark sources that occur in Vision-Flan are completely removed, including ObjectNet, Places365, ImageNet-A, ImageNet-R, VizWiz, and GQA. We additionally remove the three Vision-Flan tasks derived from A-OKVQA, although A-OKVQA belongs to the

Table 5: Composition of the representation-learning data. The "Samples" column denotes the maximum number of samples used from each subset during training.

Subset	Samples
ImageNet-1K	100,000
N24News	50,000
HatefulMemes	10,000
VOC2007	10,000
SUN397	20,000
OK-VQA	10,000
A-OKVQA	20,000
DocVQA	40,000
InfographicsVQA	25,000
ChartQA	28,000
Visual7W	70,000
VisDial	130,000
CIRR	30,000
$\text{VisualNews}_{\text{t2i}}$	100,000
$\text{VisualNews}_{\text{i2t}}$	100,000
$\text{MSCOCO}_{\text{t2i}}$	100,000
$\text{MSCOCO}_{\text{i2t}}$	120,000
NIGHTS	20,000
WebQA	20,000
MSCOCO	100,000
Vision-Flan	170,026

Table 6: Benchmark-overlapping datasets completely removed from Vision-Flan. Multiple Vision-Flan tasks derived from the same source dataset are grouped into one row.

Benchmark Dataset	MMEB Split	Removed Vision-Flan Tasks	#Removed
ObjectNet	OOD	Object Classification	1,000
Places365	OOD	Image Classification	1,000
ImageNet-A	OOD	Image Classification	1,000
ImageNet-R	OOD	Image Classification; Image Domain Classification	2,000
VizWiz	OOD	Question Answering; Image Captioning	2,000
GQA	OOD	GQA	999
A-OKVQA	IND	VQA; Rationale Generation; Answer–Rationale Matching	3,000
Total			**10,999**

IND split of MMEB, to further prevent direct overlap between training and evaluation data. Table 6 summarizes the source datasets completely removed at this stage.

The remaining MMEB OOD benchmarks, including Country211, ScienceQA, TextVQA, OVEN, FashionIQ, EDIS, Wiki-SS-NQ, and the RefCOCO family, do not appear as complete source tasks in the original Vision-Flan collection and therefore require no task-level removal.

Image-level Exact and Near-duplicate Removal. Dataset-level filtering alone is insufficient to fully eliminate evaluation leakage. Even when a training task cannot be identified as a complete benchmark source, a small fraction of its images may still coincide with, or be visually near-identical to, samples from an evaluation split. This is particularly relevant when multiple datasets reuse images from common sources such as COCO.

Therefore, after dataset-level filtering, we uniformly perform image-level decontamination against *all remaining evaluation benchmarks*, without separating representation and generation benchmarks

into independent pipelines. The evaluation pool includes the remaining MMEB benchmarks requiring sample-level auditing, together with MMMU, MMStar, MMBench, HallusionBench, MMVet, and AI2D.

For every evaluation image, we construct two complementary fingerprints: a byte-level MD5 hash and a 256-bit average perceptual hash (aHash). The former detects exact file duplicates, while the latter captures visually near-identical images caused by resizing, recompression, or minor transformations.

Given an image x, we first convert it to grayscale and resize it to 16×16 using LANCZOS interpolation. Let $\{p_i\}_{i=1}^{256}$ denote the resulting pixel intensities. Their mean intensity is

$$\bar{p} = \frac{1}{256} \sum_{i=1}^{256} p_i. \tag{11}$$

The corresponding 256-bit perceptual hash is defined as

$$h_{\mathrm{a}}(x)_i = \mathbb{I}\left[p_i \geq \bar{p}\right], \qquad i = 1, \ldots, 256, \tag{12}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

For a training image x and an evaluation image x_e, we first check exact equality using the MD5 fingerprints,

$$h_{\mathrm{MD5}}(x) = h_{\mathrm{MD5}}(x_e). \tag{13}$$

If they are not byte-identical, we further measure the Hamming distance between their perceptual hashes:

$$d_H\left(h_{\mathrm{a}}(x), h_{\mathrm{a}}(x_e)\right) = \sum_{i=1}^{256} \mathbb{I}\left[h_{\mathrm{a}}(x)_i \neq h_{\mathrm{a}}(x_e)_i\right]. \tag{14}$$

Images with a Hamming distance no greater than 5 are treated as near duplicates. Accordingly, the complete filtering criterion for a training image x is

$$\mathrm{Remove}(x) = \mathbb{I}\left[\exists x_e \in \mathcal{E},\ h_{\mathrm{MD5}}(x) = h_{\mathrm{MD5}}(x_e)\ \vee\ d_H\left(h_{\mathrm{a}}(x), h_{\mathrm{a}}(x_e)\right) \leq 5\right], \tag{15}$$

where $\mathcal{E}$ denotes the collection of all evaluation images used for decontamination.

This strategy combines *exact duplicate detection* and *perceptual near-duplicate detection*. The former reliably identifies byte-identical images, while the latter captures duplicated visual content that has undergone resizing, re-encoding, or minor image modifications.

For datasets that are retained as training sources but partially overlap with evaluation splits, only the matched samples are removed. For example, VisDial and InfographicsVQA remain in the training corpus, while individual images overlapping with their evaluation counterparts are filtered. This allows us to avoid evaluation leakage without unnecessarily discarding the entire training source.

The same image-level auditing procedure is applied to all comprehension benchmarks. During data inspection, we observed substantial overlap between the original Vision-Flan AI2D task and the AI2D evaluation data, further demonstrating that dataset-name-based filtering alone is insufficient and motivating a unified image-level decontamination stage after dataset-level removal.

Overall, Vision-Flan is reduced from 186,103 original samples to 184,031 samples after format conversion, task-aware sampling, and basic deduplication. After dataset-level benchmark removal and subsequent image-level exact and near-duplicate filtering, the final cleaned Vision-Flan subset contains 170,026 samples.

A.3 Natural-Answer Language Modeling Data

Directly using retrieval targets as language-modeling supervision is undesirable because many MMEB VQA and classification samples contain only one- or two-word positive targets. In preliminary experiments, we observed that continuously training on these extremely short responses rapidly shifted the generation behavior of the MLLM toward short and direct answers and substantially degraded performance on open-ended generation tasks.

To alleviate this issue, we construct a natural-answer supervision set, denoted as *natans*. Instead of directly using the original short retrieval targets, we employ Qwen3-VL-8B-Instruct as a teacher

Table 7: Statistics of the natural-answer annotations after quality filtering.

Subset	Generated	Retained	Retention
DocVQA	39,462	36,786	93%
InfographicsVQA	23,945	19,755	83%
ChartQA	28,000	20,877	75%
HatefulMemes	8,499	6,277	74%
OK-VQA	9,008	5,886	65%
VOC2007	7,843	4,587	58%
A-OKVQA	17,055	9,270	54%
Visual7W	69,816	34,839	50%
Vision-Flan	170,026	74,348	44%
SUN397	19,849	7,871	40%
ImageNet-1K	99,999	37,345	37%
N24News	48,987	14,823	30%
$\text{MSCOCO}_{\text{i2t}}$	19,200	19,200	100%
$\text{VisualNews}_{\text{i2t}}$	19,200	19,200	100%
Total	**580,889**	**311,064**	–

model to generate natural responses for the corresponding images and questions, and use these responses as language-modeling targets.

Before teacher inference, retrieval-specific instruction prefixes and image placeholders are removed from the original queries, such that the teacher receives natural visual questions rather than retrieval-oriented prompts. For image-to-text retrieval subsets, the input is converted into a standard image-captioning instruction. We do not impose additional reasoning templates or fixed answer formats, allowing the teacher model to retain its native response style.

We further use the same image-resolution constraints for teacher generation and student training, preventing the generated supervision from relying on visual details unavailable to the student model.

Answer Consistency Filtering. Teacher-generated responses are not automatically assumed to be correct. For VQA and classification subsets, we normalize both the generated response and the ground-truth answer and retain a sample only when one lexically contains the other. This criterion allows the teacher response to contain additional explanation while filtering responses that are inconsistent with the original supervision.

For captioning subsets, where short-answer matching is not applicable, we instead apply response-length filtering. After deduplication and consistency filtering, 311,064 natural-answer annotations are retained from 580,889 generated responses across 14 training subsets. Table 7 reports the resulting annotation statistics.

The retention rate varies substantially across task types. VQA datasets generally achieve higher retention rates, whereas classification datasets have lower coverage because a natural teacher response may be semantically correct without explicitly containing the original category label. We therefore adopt a conservative filtering strategy that prioritizes the reliability of language-modeling supervision over annotation coverage.

B Sensitivity Analysis

We investigate the sensitivity of CLUE to several key hyperparameters, including the number of learnable queries in CLE, the weight of the ECS objective λ_{ecs}, and the weight of the diversity regularization λ_{reg}. As shown in Tables 8, 9, and 10, CLUE remains relatively stable over a reasonable range of hyperparameter choices, indicating that its performance does not rely on carefully tuned configurations. Based on these results, we use 16 learnable queries, $\lambda_{\text{ecs}} = 0.1$, and $\lambda_{\text{reg}} = 0.1$ as the default configuration throughout our experiments.

Table 8: Sensitivity analysis of the number of learnable queries in CLE.

N	Generation	MMEB
1	61.1	70.5
4	61.2	70.7
8	61.2	70.9
16	61.2	**71.2**
32	61.1	71.0

Table 9: Sensitivity analysis of the ECS loss weight λ_{ecs}.

λ_{ecs}	Generation	MMEB
0.05	61.3	70.8
0.10	61.2	**71.2**
0.20	61.0	71.0

Table 10: Sensitivity analysis of the diversity regularization weight λ_{reg}.

λ_{reg}	Generation	MMEB
0.05	61.2	71.0
0.10	61.2	**71.2**
0.20	61.1	71.0

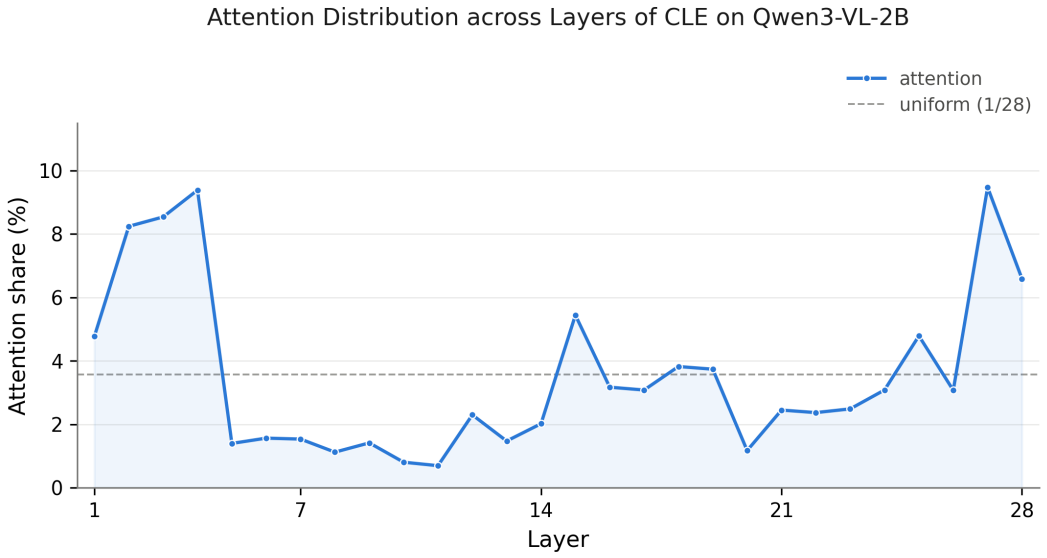


Figure 3: Visualization of cross-layer attention in CLE on Qwen3-VL-2B. The blue curve shows the average cross-attention weight assigned to the input token from each Transformer layer, while the dashed line denotes uniform attention $(1/28)$.

C VISUALIZATION OF CROSS-LAYER ATTENTION

To better understand how CLE aggregates information from different Transformer depths, we visualize the cross-attention behavior of CLE on Qwen3-VL-2B. Specifically, we average the cross-attention weights over the learnable router queries and evaluation samples, and measure the attention assigned to the layer-wise input tokens $\{h^l_{\text{last}}\}_{l=1}^{28}$. Figure 3 shows the average cross-attention weights assigned by CLE to these 28 layer-wise input tokens, where the dashed line denotes the uniform-attention baseline of $1/28$.

The learned attention exhibits a clearly non-uniform and multi-peak distribution. Several early-layer tokens, particularly those from Layers 2–4, receive substantially higher attention than the uniform baseline, while the attention decreases across much of the intermediate depth. The weights increase again for several middle-to-late layer tokens, with a prominent peak near the end of the Transformer. Notably, the attention is not concentrated solely on the final-layer token: the token from Layer 27 receives one of the highest weights, while several early and intermediate layers also contribute substantially.

This pattern suggests that CLE selectively aggregates information from multiple Transformer depths rather than simply relying on the final-layer representation or assigning equal importance to all layer-wise inputs. The heterogeneous attention allocation is consistent with the motivation of CLE that representations from different Transformer depths can provide complementary information for multimodal embedding construction. Meanwhile, the relatively high attention assigned to late-layer tokens indicates that high-level representations remain important, while the cross-layer branch additionally incorporates information from earlier and intermediate depths. Overall, the visualization provides evidence that the learnable router queries make non-trivial use of the layer-wise representations when constructing the final multimodal embedding.