# FAM: Fine-Grained Alignment Matters in Multimodal Embedding Learning with Large Vision-Language Models

**Tianhang Xiang**[1, 2*], **Yirui Li**[1*], **Lizhao Liu**[3], **Hongyan Zhi**[1], **Chuanshen Chen**[1], **Qing Du**[1†], **Mingkui Tan**[1, 2]

[1]South China University of Technology,
[2]Peng Cheng Laboratory,
[3] Tencent AI Lab
{sexiangtianhang, se_liyirui, sezhihongyan, sechenchuanshen}@mail.scut.edu.cn;
lizhaoliu@tencent.com; {duqing, mingkuitan}@scut.edu.cn

## Abstract

Learning multimodal representation is a fundamental task that supports a wide range of applications such as visual-text retrieval. While pioneering approaches *e.g.,* CLIP paves the way by learning separated encoders for different modalities, they struggle to model complex interactions between modalities, resulting in inferior vision and language representation. Recently, researchers have begun to leverage powerful Large Vision-Language Models (LVLMs) for unimodal or multimodal encoding, showing substantial improvement over separated encoder methods. However, we find that directly adapting LVLMs to embedding models suffers from insufficient visual representation and coarse multimodal alignment. To address these issues, we propose a simple yet effective Fine-grained Alignment Matters (FAM) method to achieve fine-grained vision-language embedding learning with LVLMs. First, to close the gap between the pure generation and multimodal embedding using LVLMs, we propose Multi-granularity Aligned Contrastive (MAC) to explicitly learn and align fine-grained modality representations at multiple granularity levels using image-text pairs. Second, to mitigate the insufficiency of visual representation during adapting LVLMs to downstream embedding tasks, we propose a Vision Embedding Inversion (VEIN) training strategy to encourage the extracted embeddings to preserve fine-grained visual features. Extensive experiments demonstrate the effectiveness of our method, which achieves superior performance on various downstream multimodal datasets.

**Code** — https://github.com/TianhangXiang/FAM

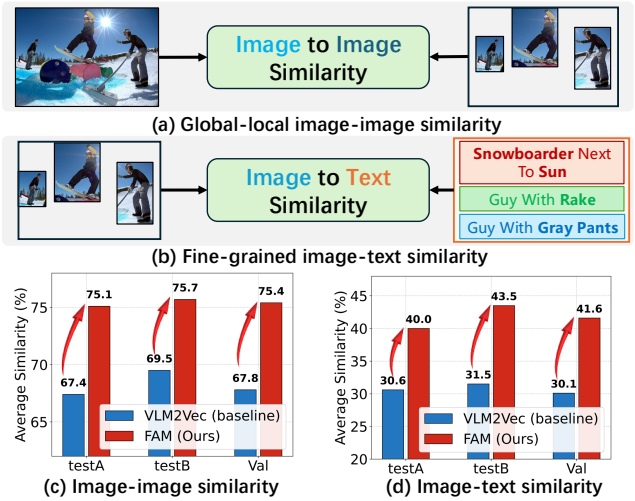



Figure 1: Illustration of the proposed tasks for evaluating multimodal alignment and fine-grained visual representation. (a) Global-local image-image similarity: evaluate the similarity between a global image embedding and object embeddings to measure the preservation of visual details; (b) Fine-grained image-text similarity: evaluate the similarity between each object embedding and its associated caption embedding within an image to assess multimodal alignment. (c) and (d) show quantitative results on the RefCOCO+ (Yu et al. 2016) dataset across `testA`, `testB` and `val` splits.

## Introduction

Multimodal representations play a crucial role in bridging data across various modalities, enabling multimodal understanding for a wide range of downstream applications, including image-text retrieval (Cao et al. 2022; Wu et al. 2021), visual question answering (Antol et al. 2015; Wu et al. 2017; Li et al. 2021), and retrieval-augmented generation (Yasunaga et al. 2022; Xia et al. 2024). Pioneering approaches, such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), employ dual-encoder architectures trained with large-scale image-text contrastive objectives, achieving notable text-image retrieval performance. However, these methods process text and images separately or perform shallow fusion of visual and textual information. Furthermore, these models exhibit limited reasoning capabilities, particularly in complex reasoning tasks.

Recently, with the rapid development and exceptional performance of large vision-language models (LVLMs) (Liu et al. 2023, 2024; Bai et al. 2025) in multimodal reasoning, researchers have begun to leverage powerful LVLMs for multimodal representation learning (Jiang et al. 2024; Lin et al. 2024; Jiang et al. 2025; Lan et al. 2025; Gu et al. 2025).

*Equal Contribution
†Corresponding Author

As a representative method, VLM2Vec (Jiang et al. 2025) constructs the Massive Multimodal Embedding Benchmark (MMEB), which reformulates four multimodal meta-tasks into embedding tasks. They adapt generative LVLMs to embedding models through contrastive learning on interleave image-text data from MMEB, showing substantial improvement over separated encoder methods. However, since LVLMs are trained in a generative paradigm focusing on high semantic text generation, which usually biases towards dominant language modality and overlooks visual representation according to prior studies (Liu, Zheng, and Chen 2024; Fu et al. 2025). Compared to text generation, embedding learning is more sensitive to fine-grained feature and modality alignment. We argue that existing methods that directly adapt generative LVLMs to representative embedding models may suffer from insufficient visual representation (*e.g.,* insufficient representation for local regions in the whole embedding) and coarse multimodal alignment, which harm the vision-language embedding learning.

To inspect the quality of visual representation and multimodal alignment, we conduct pilot studies on two similarity evaluation tasks as illustrated in Figure 1 (a)(b). To evaluate visual representation quality and multimodal alignment, we compute the average similarity between the global image and its corresponding object crops, and the average similarity between each object and its associated caption within the same image. As the results 1 (c)(d) show, the VLM2Vec is much worse than our approach in both image-image and image-text similarity evaluations, indicating the degradation of visual representation and coarse multimodal alignment.

To address these issues, we propose a simple yet effective Fine-grained Alignment Matters (FAM) method to achieve fine-grained vision-language embedding learning with LVLMs. First, before adapting LVLMs to embedding models, we propose Multi-granularity Aligned Contrastive (MAC) to mitigate the semantic gap between generative tasks and fine-grained feature-sensitive embedding tasks. Specifically, we design multiple contrastive losses from coarse-grained to fine-grained to explicitly align fine-grained features across modalities in LVLMs using image-text pairs. These objectives encourage the model to explore fine-grained feature alignment between images and their textual descriptions. Second, during the adaptation of LVLMs, we propose a **V**ision **E**mbedding **IN**version (VEIN) training strategy to encourage the extracted embeddings to preserve sufficient fine-grained visual representation. Specifically, VEIN first randomly replace a part of the visual features with mask tokens, then use the visual embedding to guide the reconstruction of masked features. Lastly, a reconstruction loss is applied between the input visual features and the reconstructed features.

In summary, our main contributions are as follows:

- To the best of our knowledge, we are the first to identify and alleviate two critical issues in directly adapting generative LVLMs to embedding models: insufficient visual representation and coarse modality alignment.
- To effectively align multimodality, we propose Multi-granularity Aligned Contrastive (MAC). Moreover, to enhance the visual representation, we also introduce Vision Embedding Inversion (VEIN) training strategy to preserve detailed visual features in learned embeddings.
- Extensive experiments on a range of multimodal datasets demonstrate the effectiveness of our approach, achieving superior performance compared to existing methods.

## Related Work

### Multimodal Representation Learning

Multimodal embeddings have emerged as a crucial area of research, aiming to unify information from various modalities, such as vision and language, into a cohesive representation space. This integration facilitates a more comprehensive understanding across different types of data. Early studies in this field (Radford et al. 2021; Jia et al. 2021; Li et al. 2022; Zhai et al. 2023) predominantly utilized dual-stream architectures, which independently encode textual and visual information. These models employ dual-encoder frameworks that leverage large-scale, weakly supervised image-text pairs, utilizing contrastive learning techniques to develop universal representations. However, these approaches often struggle with tasks involving interleaved text-image processing and complex instructions, indicating a need for more effective solutions in multimodal interactions.

### Multimodal Embeddings with LVLMs

The rapid advancement of vision-language models (VLMs) has significantly propelled multimodal embedding research by enabling unified processing and understanding of diverse input modalities. MegaPairs (Zhou et al. 2024), a large-scale dataset for multimodal instruction retrieval, has set promising results on standard benchmarks. Prior works such as E5-V (Jiang et al. 2024) and VLM2Vec (Jiang et al. 2025) have leveraged contrastive learning to adapt large vision-language models (LVLMs) into effective embedding frameworks, exploiting their capacity to seamlessly integrate interleaved text and image inputs. More recent studies (Lin et al. 2024; Liu et al. 2025) have introduced retrieval-based strategies that employ reranking models to identify top-$K$ candidates, further enhancing multimodal embedding performance. Additionally, LLaVE (Lan et al. 2025) improves representation learning for negative pairs, achieving state-of-the-art performance on the MMEB benchmark (Jiang et al. 2025) and demonstrating zero-shot generalization to text-video retrieval, underscoring its versatility across embedding tasks. However, these approaches predominantly rely on single-stage instruction tuning, which often results in suboptimal modality alignment and loss of fine-grained visual details. In contrast, UniME (Gu et al. 2025) adopts a two-stage training scheme that distills knowledge from a language expert but depends solely on unimodal supervision, limiting its ability to capture rich cross-modal semantics.

## Notations and Preliminaries

We study unified multimodal representation learning under a contrastive framework. Each training instance consists of a query $q$ and a set of candidates $\{c^+, c_1^-, \ldots, c_K^-\}$, where $c^+$

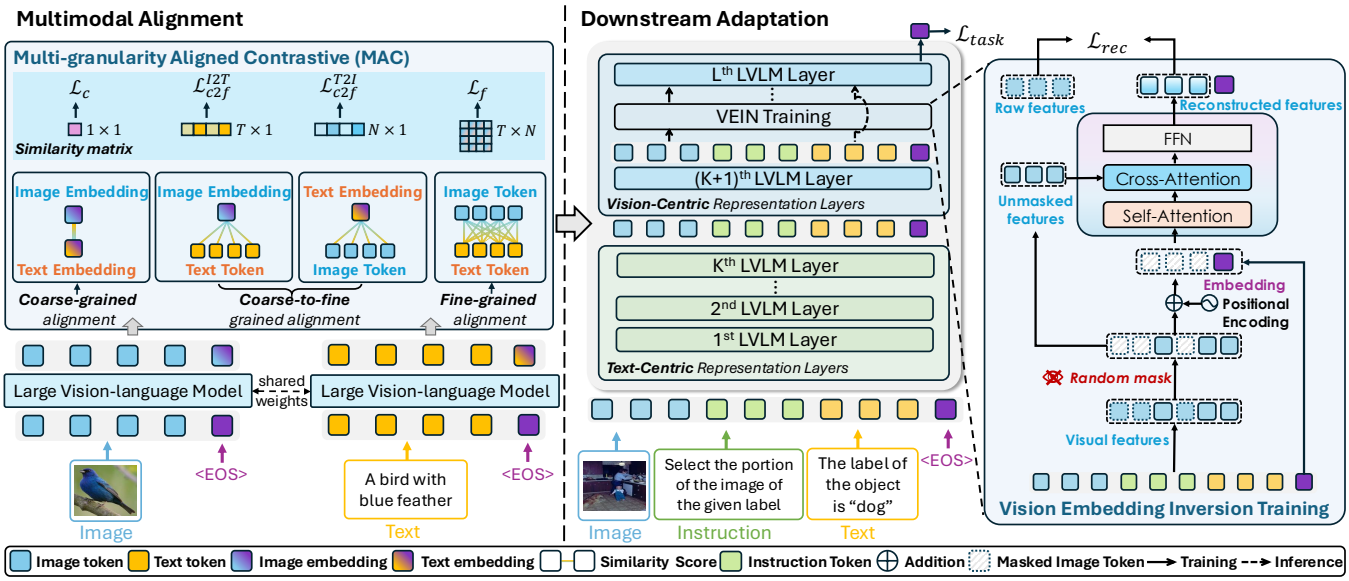



Figure 2: The overall framework of the proposed method FAM. It consists of two training stages: (1) **Multimodal Alignment**: We propose Multi-granularity Aligned Contrastive (MAC) learning to align visual and textual features at coarse, coarse-to-fine, and fine levels, leveraging multiple contrastive losses ($\mathcal{L}_{\text{c}}, \mathcal{L}_{c2f}^{\text{I2T}}, \mathcal{L}_{c2f}^{\text{T2I}}, \mathcal{L}_{f}$) for better feature alignment. (2) **Downstream Adaptation**: We propose Vision Embedding Inversion (VEIN) training strategy to preserve fine-grained visual information during adaptation. VEIN introduces a masked visual feature modeling strategy where visual features are randomly masked and reconstructed under a reconstruction loss $\mathcal{L}_{\text{rec}}$, aiming to enhance the visual representation capacity of LVLMs.

is the positive candidate and $c_i^-$ are negatives. To make all inputs compatible with LVLMs, we follow the instruction-based input format, denoted as $x_{\text{ins}}$:

$$x_{\text{ins}} = [\text{V; Instruction: } \{\textit{instruction}\}\text{; Text; <EOS>}]. \quad (1)$$

The LVLM $\mathcal{M}$, composed of $L$ stacked layers, produces hidden states $\mathbf{H}^l = \mathcal{M}^l(x_{\text{ins}})$, and the final-layer representation is obtained from the <EOS> token:

$$\mathbf{e}_{ins} = \mathcal{M}^L(x_{\text{ins}})[\text{<EOS>}]. \quad (2)$$

Given the indices of visual and textual tokens, we can further access their intermediate hidden states at any layer $l$:

$$\mathbf{H}_{\text{img}}^l = \mathbf{H}^l[\texttt{VIS\_INDEX}], \quad \mathbf{H}_{\text{txt}}^l = \mathbf{H}^l[\texttt{TXT\_INDEX}], \quad (3)$$

where $[\cdot]$ denotes the indexing operation.

## Method

In this section, we first propose an overview of the proposed FAM methods. Next, we present our Multi-granularity Aligned Contrastive (MAC). Then, we present Vision Embedding Inversion (VEIN) training strategy. Last, we summarize the overall optimization objectives.

### Overview scheme of FAM

The overall scheme of our Fine-grained Alignment Matters (FAM) is presented in Figure 2. We transform the generative LVLM into an embedding model following an align-before-adapt training paradigm. The paradigm is composed of the alignment stage and the adaptation stage. In the alignment stage, we introduce Multi-granularity Aligned Contrastive (MAC) which applies multiple contrastive loss objectives from coarse to fine to align the representations between vision and language using image-text pairs. In the adaptation stage, we adapt the aligned model to various downstream multimodal embedding tasks. To further enhance the fine-grained visual representation, we introduce a Vision Embedding Inversion (VEIN) training strategy, which reconstructs the masked visual features with the guidance of embeddings. It explicitly encourages the fine-grained visual information preservation in the extracted embeddings.

### Multi-granularity Aligned Contrastive

In the alignment stage, we aim to align the fine-grained vision-language representation using multi-granularity loss functions with image-text pair data. Given a batch of $B$ image-text pairs $\{p^b = (\text{img}^b, \text{txt}^b)\}_{b=1}^B$, we process each image and text separately. For each modality, we prepend a simple representative instruction (*e.g.,* "Represent the given image.") and append an <EOS> token before feeding it into the model $\mathcal{M}$. This process yields a set of representative embeddings $\{\mathbf{e}_{\text{img}}^b, \mathbf{e}_{\text{txt}}^b \in \mathbb{R}^d\}_{b=1}^B$ for the image and text modalities, respectively. In addition, the model outputs token-level hidden states for both modalities: $\{\mathbf{H}_{\text{img}}^b \in \mathbb{R}^{N\times d}, \mathbf{H}_{\text{txt}}^b \in \mathbb{R}^{T\times d}\}_{b=1}^B$, where $N$ and $T$ denote the number of image patches and text tokens, and $d$ is the embedding dimension. After extracting both the globals embeddings and token-level features, we apply contrastive alignment at three levels: coarse, coarse-to-fine, and fine-grained.

**Coarse-grained Contrastive Alignment** To achieve a basic vision-language alignment, we employ a contrastive learning objective on embedding from different modalities. Given the batch of representative embeddings $\{\mathbf{e}^b_{\text{img}}\}_{b=1}^B$ and $\{\mathbf{e}^b_{\text{txt}}\}_{b=1}^B$, we use the InfoNCE loss (Oord, Li, and Vinyals 2018) to align image and text embeddings at the instance level:

$$\mathcal{L}_c = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp\left(g(\mathbf{e}^b_{\text{img}}, \mathbf{e}^b_{\text{txt}})\right)}{\sum_{b'=1}^{B}\exp\left(g(\mathbf{e}^b_{\text{img}}, \mathbf{e}^{b'}_{\text{txt}})\right)}, \tag{4}$$

where $g(\mathbf{u}, \mathbf{v}) = \frac{1}{\tau}\frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}$ denotes the scaled cosine similarity with temperature parameter $\tau$.

**Coarse-to-Fine Contrastive Alignment** To encourage the model to capture fine-grained correspondences beyond instance-level alignment, we apply dual contrastive losses to align the embeddings from one modality with the token-level features from the other. We compute the similarity between the image embedding and the text hidden states as:

$$\mathbf{s}^{i,j}_{\text{I2T}} = \frac{1}{T}\sum_{t=1}^{T} g\left(\mathbf{e}^i_{\text{img}}, \mathbf{H}^j_{\text{txt}}[t]\right), \tag{5}$$

where $i, j$ represent different samples, $\mathbf{H}^j_{\text{txt}}[t]$ denotes the feature of the $t$-th text token in the $j$-th sample. Symmetrically, the similarity between the text embedding and the image hidden states is:

$$\mathbf{s}^{i,j}_{\text{T2I}} = \frac{1}{N}\sum_{n=1}^{N} g\left(\mathbf{e}^i_{\text{txt}}, \mathbf{H}^j_{\text{img}}[n]\right), \tag{6}$$

Then, the coarse-to-fine contrastive losses for both directions are defined as:

$$\mathcal{L}^{\text{I2T}}_{c2f} = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp\left(\mathbf{s}^{b,b}_{\text{I2T}}\right)}{\sum_{b'=1}^{B}\exp\left(\mathbf{s}^{b,b'}_{\text{I2T}}\right)}, \tag{7}$$

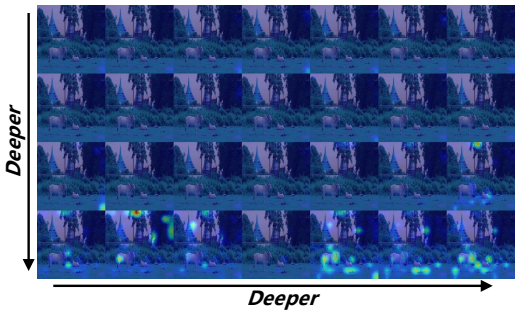
$$\mathcal{L}^{\text{T2I}}_{c2f} = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp\left(\mathbf{s}^{b,b}_{\text{T2I}}\right)}{\sum_{b'=1}^{B}\exp\left(\mathbf{s}^{b,b'}_{\text{T2I}}\right)}. \tag{8}$$

Here, $\mathbf{s}^{b,b}_{\text{I2T}}$ and $\mathbf{s}^{b,b}_{\text{T2I}}$ correspond to the similarity scores of matched image-text pairs, the final loss is computed as the average of the two directions:

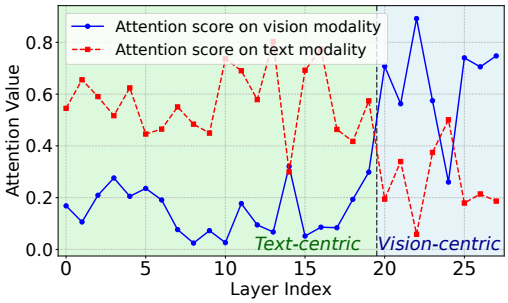
$$\mathcal{L}_{c2f} = \frac{1}{2}\left(\mathcal{L}^{\text{I2T}}_{c2f} + \mathcal{L}^{\text{T2I}}_{c2f}\right), \tag{9}$$

**Fine-grained Contrastive Alignment** To further enhance cross-modal alignment at a finer granularity, we apply an additional contrastive loss to maximize the average similarity between text tokens and image patches within a pair. Specifically, for each image-text pair in the batch, we calculate the similarity matrix $\mathbf{S}^{i,j} \in \mathbb{R}^{N\times T}$:

$$\mathbf{S}^{i,j}_{n,t} = g\left(\mathbf{H}^i_{\text{img}}[n], \mathbf{H}^j_{\text{txt}}[t]\right), \tag{10}$$




(a) Attention heatmap on visual features.




(b) Attention score on text and visual features.

Figure 3: Layer-wise attention patterns from the embedding token to all other tokens in LVLMs on the MSCOCO_i2t evaluation set, which uses instruction–image pairs for image caption retrieval, with the instruction *"Find an image caption describing the given everyday image"*.

For each image-text pair $(i, j)$, we aggregate the similarities by taking the mean over all patch-token pairs:

$$\mathbf{s}^{i,j}_{\text{f}} = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbf{S}^{i,j}_{n,t}, \tag{11}$$

The fine-grained contrastive loss is then formulated as:

$$\mathcal{L}_{\text{f}} = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp\left(\mathbf{s}^{b,b}_{\text{f}}\right)}{\sum_{b'=1}^{B}\exp\left(\mathbf{s}^{b,b'}_{\text{f}}\right)}, \tag{12}$$

## Visual-enhanced Adaptation

In the adaptation stage, we adapt the model after alignment to various downstream multimodal embedding tasks through contrastive learning on instruction-formatted inputs.

**Adaptation to Downstream Embedding Tasks** Following VLM2Vec (Jiang et al. 2025), we apply a contrastive loss on extracted embeddings between queries and targets from the downstream multimodal tasks. Given a batch of query-candidate pairs $\{(q, c)\}$, each query $q$ and candidate $c$ is converted into instruction-formatted inputs $q_{\text{ins}}$ and $c_{\text{ins}}$. The model $\mathcal{M}$ then produces embeddings $\mathbf{e}_q$ and $\mathbf{e}_c$ as shown in Equation 2. The task loss is defined as:

$$\mathcal{L}_{\text{task}} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp\left(g(\mathbf{e}^k_q, \mathbf{e}^k_{c^+})\right)}{\sum_{j=1}^{B}\exp\left(g(\mathbf{e}^k_q, \mathbf{e}^j_c)\right)}. \tag{13}$$

**Vision Embedding Inversion Training** To better preserve rich visual information in the generated embeddings, we propose the Vision Embedding Inversion (VEIN) training strategy. The core idea of VEIN is to ensure that fine-grained visual features can be decoded from the embeddings if these features are sufficiently represented. Specifically, we incorporate a module with a Transformer decoder layer $\mathcal{D}$ (Vaswani et al. 2017) into the LVLM architecture. At a designated layer $l$ of the model $\mathcal{M}$, we first extract the hidden states of the visual tokens, denoted as $\mathbf{H}^l_{\text{img}} \in \mathbb{R}^{N\times d}$, as well as the corresponding embedding $\mathbf{e}^l \in \mathbb{R}^d$, where $N$ is the number of visual tokens and $d$ is the feature dimension. To encourage the model to encode more visual information into the embeddings, we randomly mask a proportion $\gamma \in [0,1]$ of the visual tokens. Let $\mathbf{M} \in \{0,1\}^N$ be a binary mask vector, where $\mathbf{M}_j = 1$ indicates that the $j$-th token is masked. The masked visual token sequence is then constructed as follows:

$$\mathbf{H}^l_{\text{img,mask}}[j] = \begin{cases} \mathbf{m}^l, & \text{if } \mathbf{M}_j = 1 \\ \mathbf{H}^l_{\text{img}}[j], & \text{if } \mathbf{M}_j = 0 \end{cases} \tag{14}$$

where $\mathbf{m}^l \in \mathbb{R}^d$ is a learnable mask token. Next, we concatenate the embedding $e^l$ with the masked visual tokens to form the input sequence for the decoder:

$$\mathbf{X}^l = \text{concat}(\mathbf{e}^l, \mathbf{H}^l_{\text{img,mask}}) \in \mathbb{R}^{(N+1)\times d}, \tag{15}$$

and add positional encodings to obtain $\mathbf{X}^l_{\text{pos}} = \mathbf{X}^l + \text{PE}$, where PE denotes the positional encoding. The resulting sequence is then fed into the Transformer decoder layer $\mathcal{D}$, which consists of a self-attention layer, a cross-attention layer, and a feed-forward network (FFN). Formally, the process can be described as:

$$\mathbf{Z}^l = \text{SelfAttn}(\mathbf{X}^l_{\text{pos}}), \tag{16}$$

$$\mathbf{Z}^l_{\text{ca}} = \text{CrossAttn}(\mathbf{Z}^l, \mathbf{H}^l_{\text{img,unmask}}), \tag{17}$$

$$\hat{\mathbf{H}}^l_{\text{img}} = \text{FFN}(\mathbf{Z}^l_{\text{ca}}), \tag{18}$$

where $\mathbf{H}^l_{\text{img,unmask}}$ denotes the set of unmasked visual tokens, which serve as keys and values in the cross-attention layer to facilitate the reconstruction of the masked tokens. Finally, we compute the reconstruction loss only on the masked tokens, encouraging the embedding to retain sufficient visual information for accurate recovery:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathbf{M}|} \sum_{j:\mathbb{I}(\mathbf{M}_j=1)} \left[1 - \cos\left(\hat{\mathbf{H}}^l_{\text{img}}[j], \mathbf{H}^l_{\text{img}}[j]\right)\right], \tag{19}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity, $|\mathbf{M}|$ is the number of masked tokens and $\mathbb{I}(\cdot)$ is the indicator function. This process explicitly encourages the model to encode fine-grained visual information into the embeddings, as the ability to reconstruct masked visual features from the embedding serves as a strong regularization signal and improves the visual representation quality.

Where VEIN is applied within the layers of LVLMs is a critical design choice that significantly impacts the model's performance. Through qualitative and quantitative analyses, as shown in Figure 3, we observe that LVLMs extract embeddings in a way that allows their layers to be categorized into text-centric and vision-centric stages in the presentation tasks, based on the level of attention activation to the visual input. The vision-centric layers are more effective at capturing and representing visual features. To accommodate this characteristic, we conduct VEIN in the vision-centric layers. Ablation study for the choice is shown in Table 5.

## Overall Optimization

The training of FAM is conducted in two stages: alignment and adaptation. In the alignment stage, we supervise the model with three contrastive objectives to progressively enhance the vision-language alignment at different granularities: coarse-grained instance alignment, cross-modal coarse-to-fine alignment, and fine-grained token-level alignment. The total alignment loss is:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_c + \mathcal{L}_{c2f} + \mathcal{L}_f. \tag{20}$$

In the adaptation stage, we continue training the aligned model on downstream multimodal embedding tasks. The training objective here consists of the task-specific contrastive loss and the proposed VEIN-based visual reconstruction loss. The overall objective is:

$$\mathcal{L}_{\text{adapt}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{rec}}. \tag{21}$$

# Experiment

## Datasets and Metrics

In the alignment stage, we use image-caption data from LLAVA-595K (Liu et al. 2023) dataset and the caption used is the BLIP-2 (Li et al. 2023) refine version. In the adaptation stage, we follow VLM2Vec (Jiang et al. 2025) to train the model on 20 in-distribution datasets from MMEB (Jiang et al. 2025). These datasets contain four meta-tasks: Classification (Cls.), Visual Question Answering (VQA), Multimodal Retrieval (Re.) and Visual Grounding (VG), which are reforming into a retrieval task format. We report the Precision@1 on each meta-task and the overall average accuracy on both in-distribution and out-of-distribution datasets.

## Implementation Details

Following VLM2Vec, we adopt Qwen2-VL (2B and 7B) (Wang et al. 2024) and Phi-3.5-V (Abdin et al. 2024) as our backbone models. We leverage LoRA (Hu et al. 2022) with bfloat16 precision for parameter-efficient training. During the alignment stage, a LoRA module is trained and subsequently merged into the base model. In this phase, input images are resized to $672 \times 672$ with a batch size of 512. For the adaptation stage, an additional LoRA module is trained on top of the merged model, with images resized to $336 \times 336$ and a batch size of 256. VEIN is applied exclusively during the adaptation stage and incurs no additional inference overhead. It is applied to queries and targets containing image inputs, except for VQA tasks. For both stages, the LoRA rank is set to 8 and the learning rate is $2 \times 10^{-5}$. Specifically, for Qwen2-VL-2B, VEIN is applied to the last three layers, whereas for Qwen2-VL-7B and Phi-3.5-V, it is applied only to the final layer. All experiments are conducted on NVIDIA A100 GPUs (80GB).

| Method | Backbone | Per Meta-Task Score | | | | Average Score | | |
|---|---|---|---|---|---|---|---|---|
| | | Classification | VQA | Retrieval | VG | IND | OOD | Overall |
| # Datasets Numbers | | 10 | 10 | 12 | 4 | 20 | 16 | 36 |
| *Dual-Encoder-based Models (Zero-shot)* | | | | | | | | |
| CLIP (Radford et al. 2021) | - | 42.8 | 9.1 | 53.0 | 51.8 | 37.1 | 38.7 | 37.8 |
| BLIP2 (Li et al. 2023) | - | 27.0 | 4.2 | 33.9 | 47.0 | 25.3 | 25.1 | 25.2 |
| SigLIP (Zhai et al. 2023) | - | 40.3 | 8.4 | 31.6 | 59.5 | 32.3 | 38.0 | 34.8 |
| OpenCLIP (Cherti et al. 2023) | - | 47.8 | 10.9 | 52.3 | 53.3 | 39.3 | 40.2 | 39.7 |
| UniIR (BLIP_FF) (Wei et al. 2024) | - | 42.1 | 15.0 | 60.1 | 62.2 | 44.7 | 40.4 | 42.8 |
| UniIR (CLIP_SF) (Wei et al. 2024) | - | 44.3 | 16.2 | 61.8 | 65.3 | 47.1 | 41.7 | 44.7 |
| Magiclens (Zhang et al. 2024) | | 38.8 | 8.3 | 35.4 | 26.0 | 31.0 | 23.7 | 27.8 |
| *Dual-Encoder-based Models (Fine-tuning on MMEB Training)* | | | | | | | | |
| CLIP-FFT (Radford et al. 2021) | - | 55.2 | 19.7 | 53.2 | 62.2 | 47.6 | 42.8 | 45.4 |
| OpenCLIP-FFT (Cherti et al. 2023) | - | 56.0 | 21.9 | 55.4 | 64.1 | 50.5 | 43.1 | 47.2 |
| *LVLM-based Models* | | | | | | | | |
| E5-V (Jiang et al. 2024) | LLaVA-NeXT-8B | 21.8 | 4.9 | 11.5 | 19.0 | 14.9 | 11.5 | 13.3 |
| VLM2Vec* (Jiang et al. 2025) | Qwen-2VL-2B | 57.8 | 40.9 | 60.6 | 67.1 | 59.8 | 49.2 | 55.0 |
| FAM (Ours) | Qwen-2VL-2B | **58.6** | **42.2** | **64.1** | **70.9** | **61.7** | **51.8** | **57.3** |
| VLM2Vec* (Jiang et al. 2025) | Phi-3.5-V | 52.7 | 50.8 | 59.3 | 81.0 | 63.2 | 50.4 | 57.5 |
| FAM (Ours) | Phi-3.5-V | **53.8** | **50.9** | **61.2** | **83.1** | **64.8** | **51.1** | **58.7** |
| VLM2Vec* (Jiang et al. 2025) | Qwen-2VL-7B | 61.3 | **47.8** | 65.5 | 75.8 | 65.6 | 54.2 | 60.5 |
| FAM (Ours) | Qwen-2VL-7B | **62.1** | 47.5 | **68.0** | **78.9** | **66.5** | **56.1** | **61.9** |

Table 1: Comparisons with state-of-the-art methods on MMEB benchmark. $*$ denotes results based on our reimplementation under the low-resolution ($336 \times 336$) setting for fair comparison. For Phi-3.5-V–based models, 4 crops are used for each image.

| Method | MAC | VEIN | Meta-Task Avg. Score | | | | Avg. |
|---|---|---|---|---|---|---|---|
| | | | Cls. | VQA | Re. | VG | |
| VLM2Vec | - | - | 57.8 | 40.9 | 60.6 | 67.1 | 55.0 |
| FAM (Ours) | ✓ | ✗ | 58.5 | 39.8 | 64.0 | 70.3 | 56.4 |
| FAM (Ours) | ✗ | ✓ | 58.4 | **42.9** | 60.2 | 69.2 | 55.9 |
| FAM (Ours) | ✓ | ✓ | **58.6** | 42.2 | **64.1** | **70.9** | **57.3** |

Table 2: Ablation studies on the proposed MAC and VEIN.

## Comparison with State-of-the-arts

Table 1 provides a comprehensive comparison on the MMEB benchmark, evaluating our method against dual-encoder models (*e.g.,* CLIP (Radford et al. 2021), OpenCLIP (Cherti et al. 2023)) and recent LVLM-based models. Benefiting from powerful language understanding and multimodal reasoning ability, LVLM-based models generally outperform dual-encoder methods. FAM surpasses VLM2Vec by a noticeable margin with +2.3, +1.4 and +1.2 on the Qwen2-VL (2B and 7B) and Phi-3.5-V backbones, respectively. Notably, FAM delivers significant improvements on the multimodal retrieval task and the visually sensitive grounding task, demonstrating its superior fine-grained cross-modal alignment and visual representations.

## Ablation Studies

**Ablation study on the proposed method** We conduct comprehensive ablation studies to assess the effectiveness of the proposed MAC and VEIN modules, as shown in Table 2. Removing either component leads to a performance drop, indicating that both are beneficial. Compared to the baseline VLM2Vec, adding MAC or VEIN individually improves the average score to 55.9 and 56.4, respectively. Combining both modules yields the best performance of 57.3, with notable gains in multimodal retrieval and visual grounding tasks, confirming the effectiveness of proposed MAC and VEIN.

**Ablation study on training strategy** As MAC involves additional training data, we perform experiments to ensure that the observed improvements primarily originate from our training strategy rather than the increased data volume. To enable a fair comparison, we transform the extra image-text pairs used in MAC into retrieval task data and add them to the VLM2Vec training dataset. As shown in Table 3, adding alignment data slightly improves the average score to 55.3, suggesting that simply increasing the training data provides limited gains. In contrast, our training strategy efficiently utilizes the image-text pairs data to achieve fine-grained modality alignment and achieves a notable improvement.

**Analysis on Multi-granularity Aligned Contrastive** Since MAC leverages multiple contrastive losses to align modality features at different granularities, we perform an ablation by progressively adding these losses. We evaluate their impact on both the alignment-only model and the full training pipeline. As shown in Table 4, incorporating finer-grained losses consistently improves performance in both settings, confirming their effectiveness in enhancing multi-granularity alignment and downstream task performance.

| Method | Data | MAC | Meta-Task Avg. Score | | | | Avg. |
|---|---|---|---|---|---|---|---|
| | | | Cls. | VQA | Re. | VG | |
| VLM2Vec | D. | ✗ | 57.8 | 40.9 | 60.6 | 67.1 | 55.0 |
| VLM2Vec | A.+D. | ✗ | 58.5 | **41.6** | 59.6 | 68.9 | 55.3 |
| FAM (Ours) | A.+D. | ✓ | **58.5** | 39.8 | **64.0** | **70.3** | **56.4** |

Table 3: Ablation studies on training strategy and MAC. "A." indicates the LLAVA pretraining data for alignment and "D." indicates the training data from MMEB for adaptation.

| Stage | $\mathcal{L}_c$ | $\mathcal{L}_{c2f}$ | $\mathcal{L}_f$ | Meta-Task Avg. Score | | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | | | | Cls. | VQA | Re. | VG | |
| Align | ✓ | ✗ | ✗ | 32.6 | 7.6 | 50.2 | **55.2** | 34.0 |
| Align | ✓ | ✓ | ✗ | 35.6 | **9.1** | 51.2 | 52.1 | 35.3 |
| Align | ✓ | ✓ | ✓ | **35.9** | 8.9 | **51.4** | 52.6 | **35.4** |
| Align + Adapt | ✓ | ✗ | ✗ | 57.1 | **40.1** | 63.2 | 69.6 | 55.8 |
| Align + Adapt | ✓ | ✓ | ✗ | 58.0 | 39.8 | 63.4 | **70.4** | 56.1 |
| Align + Adapt | ✓ | ✓ | ✓ | **58.5** | 39.8 | **64.0** | 70.3 | **56.4** |

Table 4: Effect of different granularity contrastive losses on the MMEB benchmark. "Align" indicates the alignment stage, "Adapt" stands for the adaptation stage.

**Effect of Mask Ratio ($\gamma$) of VEIN** We investigate the impact of different mask ratios for VEIN during adaptation. Figure 4 shows that performance improves as the mask ratio increases at first and peaks around 0.3. After that, the improvement decreases and ratios above 0.8 make the model worse than the baseline. This trend differs from MAE (He et al. 2022), where higher mask ratios often yield better results. We attribute this discrepancy to the fact that VEIN reconstructs feature-level representations, which are more semantic and structured than the pixel-level inputs reconstructed in MAE. Excessive feature-level masking disrupts meaningful visual reconstruction and reduces effectiveness.

**Impact of VEIN Insert Positions** As shown in Figure 3, different LVLM layers attend to different modality inputs. Since VEIN is a visual reconstruction strategy, we investigate its effectiveness when conducted at different model depths. As shown in Table 5, conducting VEIN at all layers harms performance. In contrast, conducting VEIN at deeper and visually focused layers consistently improves results, with the best performance obtained when inserted into the last three layers. This confirms that placing VEIN in vision-centric layers is more beneficial than vinilla insertion that performs masked visual feature reconstruction in all layers.

## Qualitative Results

Figure 5 provides qualitative examples of retrieval and visual grounding tasks on the MMEB benchmark[1]. After applying MAC to the baseline method VLM2Vec, the model is more effective in capturing finer-grained vision–language correspondences, enabling it to retrieve images that better match complex captions with multiple objects (*e.g.,* cut pizza, salad and soda). Incorporating VEIN further enhances the representation of fine-grained visual details, allowing the model to capture subtle but crucial cues (*e.g.,* the color and shape of the car), which ultimately leads to more accurate localization of the target object.

[1]Unlike conventional visual grounding, the MMEB benchmark crops targets from multiple images rather than from the query image only, so the top match one may not originate from the same query image.

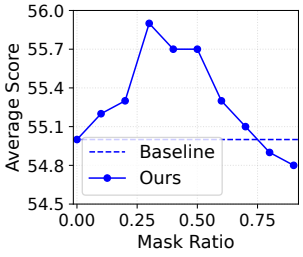



Figure 4: Effect of mask ratios. A moderate ratio (*e.g.,* 0.3) yields the best accuracy.

| VEIN Conduct Position | Avg. |
|---|---|
| N/A | 55.0 |
| All Layers | 54.2 |
| Last Layer | 55.3 |
| Second to Last Layer | 55.6 |
| Third to Last Layer | 55.2 |
| Last Three Layers | **55.9** |

Table 5: Ablation study on positions conducing VEIN. Conducting VEIN in deeper layers yields competitive or better overall performance.

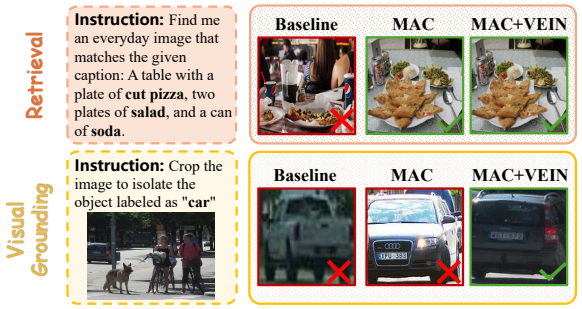



Figure 5: Qualitative examples on retrieval and visual grounding tasks after applying MAC and VEIN to the baseline method VLM2Vec on the MMEB benchmark.

## Conclusion

In this work, we investigate the limitations of directly adapting generative Large Vision-Language Models (LVLMs) for multimodal embedding learning and indentify insufficient visual representation and coarse modality alignment as key challenges. To address these issues, we propose Fine-grained Alignment Matters (FAM), a simple yet effective framework that leverages fine-grained multi-level contrastive alignment Multi-granularity Aligned Contrastive (MAC) and Vision Embedding Inversion (VEIN) training strategy to enhance fine-grained visual perseveration. Our approach explicitly aligns visual and textual features at multiple granularities and encourages the retention of detailed visual information in the learned embeddings. Extensive experiments across diverse multimodal datasets demonstrate the effectiveness of our method. In the future, we will further explore the proposed FAM on more diverse and higher-resolution settings.

## Acknowledgments

This work was partially supported by the Joint Funds of the National Natural Science Foundation of China (Grant No.U24A20327).

## References

Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H. H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H. S.; Benhaim, A.; Bilenko, M.; Bjorck, J.; Bubeck, S.; Cai, M.; Mendes, C. C. T.; Chen, W.; Chaudhary, V.; Chopra, P.; Giorno, A. D.; de Rosa, G.; Dixon, M.; Eldan, R.; Iter, D.; Goswami, A.; Gunasekar, S.; Haider, E.; Hao, J.; Hewett, R. J.; Huynh, J.; Javaheripi, M.; Jin, X.; Kauffmann, P.; Karampatziakis, N.; Kim, D.; Kim, Y. J.; Khademi, M.; Kurilenko, L.; Lee, J. R.; Lee, Y. T.; Li, Y.; Liang, C.; Liu, W.; Lin, E.; Lin, Z.; Madan, P.; Mitra, A.; Modi, H.; Nguyen, A.; Norick, B.; Patra, B.; Perez-Becker, D.; Portet, T.; Pryzant, R.; Qin, H.; Radmilac, M.; Ren, L.; Rosset, C.; Roy, S.; Saarikivi, O.; Saied, A.; Salim, A.; Santacroce, M.; Shah, S.; Shang, N.; Sharma, H.; Song, X.; Ruwase, O.; Vaddamanu, P.; Wang, X.; Ward, R.; Wang, G.; Witte, P. A.; Wyatt, M.; Xu, C.; Xu, J.; Yadav, S.; Yang, F.; Yang, Z.; Yu, D.; Zhang, C.-Y.; Zhang, C.; Zhang, J.; Zhang, L. L.; Zhang, Y.; Zhang, Y.; Zhou, X.; and Yang, Y. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *ArXiv*, abs/2404.14219.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cao, M.; Li, S.; Li, J.; Nie, L.; and Zhang, M. 2022. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Fu, S.; Bonnen, T.; Guillory, D.; and Darrell, T. 2025. Hidden in plain sight: VLMs overlook their visual representations. *arXiv preprint arXiv:2506.08008*.

Gu, T.; Yang, K.; Feng, Z.; Wang, X.; Zhang, Y.; Long, D.; Chen, Y.; Cai, W.; and Deng, J. 2025. Breaking the Modality Barrier: Universal Embedding Learning with Multimodal LLMs. *arXiv preprint arXiv:2504.17432*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jiang, T.; Song, M.; Zhang, Z.; Huang, H.; Deng, W.; Sun, F.; Zhang, Q.; Wang, D.; and Zhuang, F. 2024. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Jiang, Z.; Meng, R.; Yang, X.; Yavuz, S.; Zhou, Y.; and Chen, W. 2025. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *ICLR*.

Lan, Z.; Niu, L.; Meng, F.; Zhou, J.; and Su, J. 2025. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, L.; Lei, J.; Gan, Z.; and Liu, J. 2021. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2042–2051.

Lin, S.-C.; Lee, C.; Shoeybi, M.; Lin, J.; Catanzaro, B.; and Ping, W. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, S.; Zheng, K.; and Chen, W. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, 125–140. Springer.

Liu, Y.; Zhang, Y.; Cai, J.; Jiang, X.; Hu, Y.; Yao, J.; Wang, Y.; and Xie, W. 2025. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4015–4025.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wei, C.; Chen, Y.; Chen, H.; Hu, H.; Zhang, G.; Fu, J.; Ritter, A.; and Chen, W. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, 387–404. Springer.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11307–11317.

Wu, Q.; Teney, D.; Wang, P.; Shen, C.; Dick, A.; and Van Den Hengel, A. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163: 21–40.

Xia, P.; Zhu, K.; Li, H.; Wang, T.; Shi, W.; Wang, S.; Zhang, L.; Zou, J.; and Yao, H. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Yasunaga, M.; Aghajanyan, A.; Shi, W.; James, R.; Leskovec, J.; Liang, P.; Lewis, M.; Zettlemoyer, L.; and Yih, W.-t. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European conference on computer vision*, 69–85. Springer.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, K.; Luan, Y.; Hu, H.; Lee, K.; Qiao, S.; Chen, W.; Su, Y.; and Chang, M.-W. 2024. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*.

Zhou, J.; Liu, Z.; Liu, Z.; Xiao, S.; Wang, Y.; Zhao, B.; Zhang, C. J.; Lian, D.; and Xiong, Y. 2024. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*.