Lenses: Toward Polysemous Vision–Language Understanding

Hani Alomari
Virginia Tech
hani@vt.edu

Ali Asgarov
Virginia Tech
aliasgarov@vt.edu

Chris Thomas
Virginia Tech
chris@cs.vt.edu

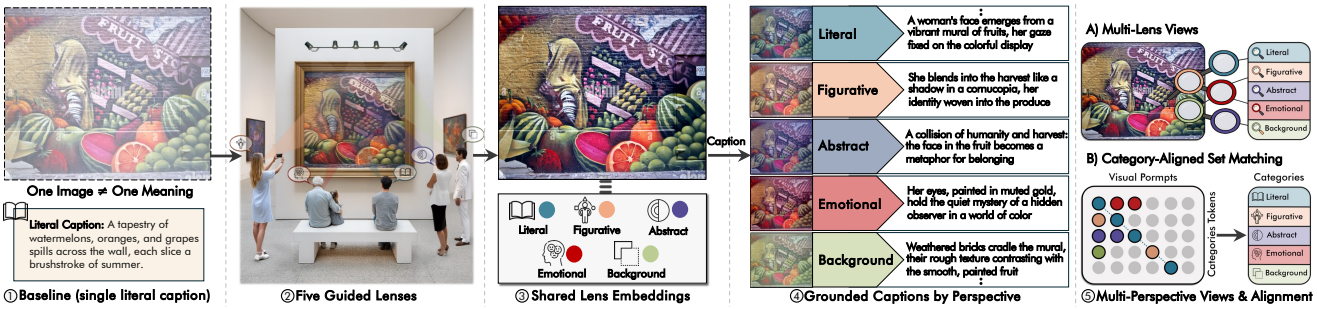

Figure 1. Overview of Lenses. (Left) Standard datasets treat each image as having one literal meaning. (Middle) Lenses decomposes images into five interpretive lenses: Literal, Figurative, Abstract, Background, and Emotional, with diverse captions for each perspective. Captions are encoded into lens-specific embedding slots. (Right) During retrieval, category-aligned set matching ensures only same-lens slots interact, enabling perspective-aware image-text matching.

Abstract

Most vision-language models assume images have a single literal meaning, even though images are inherently polysemous. We propose a retrieval paradigm that models many-to-many relationships between images and text using interpretive lenses and introduce Lenses, a multi-prompt embedding model and dataset for polysemous image-text retrieval. The Lenses dataset contains $105,669$ *images and* $732,405$ *captions, with each image paired with multiple captions and image-side prompts annotated across five categories: Literal, Figurative, Abstract, Background, and Emotional. Building on a multimodal large language model, the Lenses model uses learned lens tokens to extract lens-specific embeddings for every image and caption and compares these using a lens-masking similarity function with a global fallback that prioritizes same-lens matches while retaining a global pathway. Training uses a category-aware multi-positive contrastive loss and intra-set diversity regularization to align corresponding perspectives while preventing semantic collapse across lenses. We further propose lens-aware evaluation protocols, including category-aware ranking, that better reflect how humans match images and text. Experiments on the Lenses dataset and public benchmarks show that our model outperforms baselines on literal and non-literal retrieval and reduces over-reliance on literal cues.*

1. Introduction

People can see the same image from diverse perspectives and contexts, with different emotions that shape what they notice and how they describe it [1, 7]. For example, an image of a rotten apple among fresh ones can be described literally as 'decaying fruit' or figuratively by the well-known idiom 'one bad apple spoils the barrel'. Despite rapid progress in vision-language models (VLMs), mainstream training and evaluation (*e.g.* COCO [25], Flickr30k [32]) still treat alignment as a single-meaning, literal image–caption matching task, missing the richness of alternative interpretations. However, prior work on non-literal language in image descriptions [1, 7, 34, 45] shows that captions can include meanings that extend beyond the image's visible content. These may involve one or more aspects, for example: figurative, abstract, background, or emotional. Such meanings lead to many-to-many, non-literal alignments. This mismatch causes models to overfit literal details while underrepresenting diverse meanings. We address this with Lenses, a multi-prompt embedding model that uses these lenses to align images and captions for polysemous tasks. This extends prior research on literal versus figurative language [8], emotional image responses [7], and abstract/background dynamics in captions and art [3].

Most current VLMs learn a single global embedding per modality and optimize a contrastive objective over large

batches. This design scales and works remarkably well for *Literal* retrieval [13, 33, 47]. However, compressing an image or caption into one vector blends all its different meanings together, making the system fragile when faced with queries with multiple interpretations. Cross-attention models [20] enable fine-grained token-to-region matching but incur substantial inference cost, which limits scalability. Late-interaction models [16] defer similarity computation to query time, preserving token-level granularity while maintaining many dual-encoder advantages. Multi-embedding dual-encoder approaches (*e.g.* PVSE [36], PCME [5], SetDiv [18], MaxMatch [2]) promote diversity by learning multiple embeddings, but they typically re-aggregate them into a global representation to reduce interference. Meanwhile, VLM retrievers (BLIP/BLIP-2 [21, 22], LLaVA [26], Emu2 [37], MM-Embed [24]) produce strong universal features but still tend to pool into a single representation or attend nearly uniformly over visual tokens.

We formulate *polysemous retrieval* using lens-aware *sets of embeddings*. We introduce Lenses, a dataset that evaluates *literal* () and non-literal matching across *figurative* (), *abstract* (), *background* (), and *emotional* () lenses. Each image includes category-labeled captions and image-side prompts for all five lenses. To solve this problem, we propose a method that uses a set of special tokens to extract multiple, lens-specific embeddings for each image and caption. We compare image and caption slot sets with a category-aware similarity that leverages lens category matching as guidance; when no same-lens match exists, we fall back to a global embedding. Training uses a *multi-positive contrastive* objective to pull positive matches together and avoid treating other positives as negatives. To preserve slot diversity, we add a lens-conditioned caption–prompt alignment loss and an intra-image prompt-diversity regularizer that keeps slots for the same image distinct.

We evaluate the proposed approach on our new Lenses dataset and compare it with recent vision–language retrieval models and multi-embedding baselines, showing clear gains across both literal and non-literal lenses. In summary, our main contributions are as follows:

- We introduce a dataset of (105,669 images, 732,405 captions) that explicitly benchmarks image–text matching across five interpretive lenses: *literal*, *figurative*, *abstract*, *background*, and *emotional*.
- We propose a **multi-prompt embedding model** that uses special lens tokens to extract distinct, prompt-conditioned embeddings for both image and text, and introduce a category-aware lens-matching similarity with a global fallback for robust cross-lens matching.
- We design a **multi-positive training objective** and lens-conditioned regularizers that preserve slot diversity, leading to substantial improvements in both lens-specific and overall retrieval performance on Lenses and ArtEmis.

2. Related work

Cross-modal retrieval typically uses dual encoders that map images and text into a shared space, trained with contrastive objectives [33]. Later work improves similarity functions [18, 44], losses [5, 38], and model architectures [9, 17] to capture finer semantics.

Vision-language embedding models build directly on the shift toward semantically richer cross-modal encoders. A first line of work leverages large-scale image-text pretraining to obtain generic joint embeddings, for example BLIP [21] and BLIP-2 [22], which are trained with captioning or VQA objectives and then reused for retrieval [21, 22]. Instruction-tuned multimodal large language models such as LLaVA and related generative MLLMs adapt these encoders so that the vision or text branch becomes more query aware and better aligned with downstream semantic intents [26, 37]. E5-V [15] pushes this idea and treats an MLLM as a universal embedding function for both unimodal and multimodal inputs, trained with instruction-style prompts and contrastive objectives. UniIR [43] jointly instruction-tunes CLIP- and BLIP-based models so that a single retriever can follow natural-language task instructions across many multimodal retrieval datasets. Open-source retrievers such as MegaPairs [49], InternVL [4], and Jina [10] follow similar pipelines, fine-tuning powerful VLM backbones as bi-encoders for large-scale retrieval. However, these methods either map each input to a single pooled embedding or perform only coarse token matching and they are trained and evaluated using standard Recall@k on predominantly literal corpora.

Multi-embedding representations handle ambiguity and one-to-many alignments by moving beyond a global embedding and representing each modality with multiple embeddings. PVSE [36] learns a fixed set of semantic heads per sample and regularizes them to encourage diversity. PCME [5] models each image or caption as a mixture of Gaussian embeddings to capture semantic uncertainty. Set-based models with smooth-Chamfer objectives aggregate soft set-to-set correspondences between multiple slots [18], and MaxMatch [2] introduces a maximal one-to-one matching loss that guards against a single pair dominating under sparse supervision. Late-interaction retrievers like ColBERT encode queries and documents with multiple token vectors and score them via MaxSim [16]; ColBERT-X [28] and XTR [19] show that multi-vector retrieval scales well and extends cross-lingually while preserving fine-grained signals. Similar ideas appear in PolyViT [23], which co-trains a transformer across modalities, and DELF [29], which uses attention to select local features for retrieval.

Despite using multiple embeddings, these approaches treat them as latent capacity for generic semantic variation rather than as a *semantically diverse* set of representations. At both training and inference time, the multiple embeddings are collapsed into a single global similarity score, through set

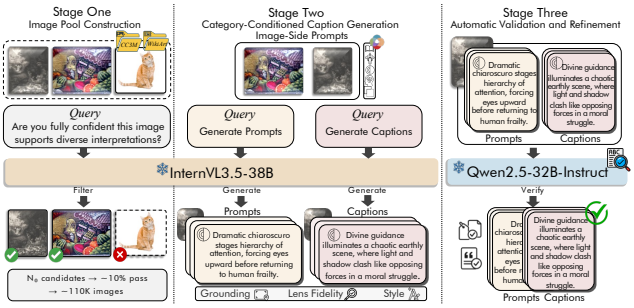

Figure 2. **Lenses construction pipeline.** (1) We filter CC3M and WikiArt using frozen InternVL3.5-38B to identify images that admit non-literal/background readings. (2) For each selected image, InternVL3.5-38B generates a set of lens conditioned texts *Literal*, *Figurative*, *Abstract*, *Background*, and *Emotional* covering both prompts and captions. (3) Qwen2.5–32B–Instruct verifies grounding and lens fidelity, discarding low-quality samples. Full generation and verification prompts are provided in the supplementary material.

pooling, mixture likelihood, or MaxSim over tokens. Supervision is drawn almost entirely from literal datasets [25, 32] and evaluated on mostly literal connections. In practice, the resulting multi-vector representations tend to encode coarse semantic factors or local detail, not systematically distinct and diverse readings of the same image. In contrast, we reorient the multi-embedding paradigm toward explicitly polysemous understanding. We introduce lens tokens and a category-aware matching objective that tie each slot to an interpretive lens, and we train and evaluate under non-literal supervision on Lenses. This yields lens-specific embeddings for each image and caption and forces the model to decide, for a given pair, which lens should drive the match.

3. Methodology

3.1. Lenses dataset construction

Large-scale collections of image-text pairs underpin modern vision-language models, but most available pairs provide only a single, predominantly literal description per image [33]. Perspective-aware supervision, where a model is encouraged to recognize that the same image can support both literal and non-literal interpretations, has been shown to improve retrieval robustness and diversity [2, 18, 36, 48]. Yet multi-caption datasets such as COCO [25] and Flickr30K [32] largely contain captions that are stylistically similar and focused on the same salient objects or actions [40, 41]. Multi-embedding approaches (*e.g.* PVSE [36], SetDiv [18], MaxMatch [2]) were therefore trained on captions that differ mainly in wording rather than meaning, and are not designed to model genuinely polysemous semantics. At the same time, collecting dense human multi-perspective annotations at scale is prohibitively expensive. Our automatic construction pipeline, summarized in Fig. 2, aims to strike a practical

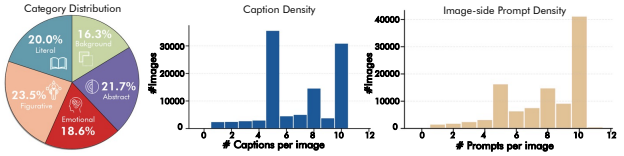

Figure 3. **Lenses label balance and annotation density.** Left: distribution of captions across the five lenses. Middle: captions per image. Right: image-side prompts per image. The dataset is roughly balanced across lenses, and most images include multiple captions and multiple prompts

Split	# Images	# Captions					
Train	99,069	686,469	137,034	161,126	148,652	112,111	127,546
Valid	1,100	7,677	1,526	1,799	1,654	1,269	1,429
Test	5,500	38,259	7,613	8,950	8,262	6,306	7,128
Total	105,669	732,405	146,173	171,875	158,568	119,686	136,103

Table 1. Dataset statistics by split and lens. Each split contains captions annotated with five perspective categories (Literal, Figurative, Abstract, Background, Emotional).

balance between scalability and label quality.

To address this gap, we construct Lenses, a large-scale multi-perspective image-text dataset in which each image is paired with category-labeled descriptions spanning five interpretive lenses: *literal*, *figurative*, *abstract*, *background*, and *emotional*. Formally, each instance comprises an image $\mathcal{I}$ and a set of captions $\mathcal{S}(\mathcal{I}) = \{(t_k, c_k)\}_{k=1}^{K}$ with $c_k \in \mathcal{C}$, where $\mathcal{C} = \{$, , , , $\}$. In addition, we construct an image-conditioned prompt set $\mathcal{P}(\mathcal{I}) = \{(p_z, c_z)\}_{z=1}^{Z}$ consisting of short, category-labeled phrases that act as compact perspective cues. Table 1 and Figure 3 summarize the label distribution and caption and prompt densities; Lenses contains 105,669 images and 732,405 captions with all five lenses represented across splits.

Image pool construction. We source candidate images from CC3M [35] and WikiArt [27, 31] to ensure coverage of both everyday scenes and artwork that naturally afford diverse interpretations beyond literal object naming. To identify images with rich interpretive potential, we filter all candidates using InternVL3.5-38B with a binary classification prompt that asks whether the image supports diverse non-literal or background readings. Approximately 10% of candidates pass this filtering stage, yielding 105,669 diverse images spanning everyday scenes, portraits, landscapes, and abstract compositions. We remove near-duplicates through perceptual hashing and CLIP-space clustering. Full prompt text and thresholds are provided in the supplementary material.

Category-conditioned caption generation. For each image $\mathcal{I}$, we generate multiple captions across the five perspective categories: *literal*, *figurative*, *abstract*, and *background*, and *emotional*. We treat large vision-language models as structured annotators: InternVL3.5-38B

is prompted with category-specific instructions that specify desired coverage, visual grounding, and style. The instructions enforce three constraints: *grounding*, each caption must be justified by visual evidence in the image; *lens fidelity*, the caption must realize the intended perspective (for example, idiomatic figurative language for the figurative category, affective language for the emotional category); and *stylistic appropriateness*, captions should be concise, non-story-like descriptions suitable for retrieval. The prompt explicitly requests at least one caption per category when possible and allows additional captions when the model identifies distinct readings. For non-literal lenses, we provide a curated phrase bank of idioms and metaphors but require that any borrowed expressions remain visually supported. We sample multiple candidates per lens and retain those that satisfy the automatic validation checks described below, forming the final caption set $\mathcal{S}(\mathcal{I})$. The exact prompt templates and phrase bank are included in the supplementary material.

Image-side prompts. For each image $\mathcal{I}$, we also generate a variable-sized set of short, category-labeled prompts that act as perspective cues, directing attention toward specific interpretive lenses. Using InternVL3.5-38B with carefully designed instructions, we ask the model to produce concise prompts that target a single lens $c_z \in \mathcal{C}$, foreground a salient visual element (object, region, spatial relation, texture, color palette, compositional cue, or thematic content), and avoid generic or lens-agnostic phrasing. Each prompt is self-contained. To encourage diversity, we vary syntactic patterns (imperatives, descriptive phrases, contrastive clauses, rhetorical questions) and discourage near-duplicate paraphrases via n-gram overlap checks. These prompts are included in the dataset and are used by our retrieval model as lens-specific query anchors. As shown in Fig. 3 (right), most images are associated with multiple prompts.

Validation and refinement. Because captions and prompts are produced automatically, we apply a second-stage verification process using Qwen2.5–32B–Instruct to validate each caption (t_k, c_k) and prompt (p_z, c_z), ensuring that texts are well formed, correctly aligned with their assigned categories, and maintain the intended perspective. Items failing any validation criterion are either revised or discarded. Finally, we apply an additional refinement stage to the test set using InternVL3.5-38B with a verification-only prompt: for each test instance, InternVL3.5-38B receives the image and all generated prompts or captions and is asked to verify visual grounding. Prompts or captions lacking sufficient grounding in the visual content are removed or revised, ensuring the test set maintains high annotation quality and consistency. We perform a small-scale human validation study on a stratified subset of test captions and prompts to assess visual grounding and lens-label fidelity; details and per-lens agreement statistics are reported in the supplementary material.

3.2. Lenses Model

We propose Lenses, a series of models designed for multi-lens retrieval based on a pre-trained VLM. It incorporates a visual encoder, typically a vision transformer [6], into an LLM [14], so that image tokens are directly processed by the LLM. The resulting VLM can handle diverse multimodal inputs by converting any input into a sequence of tokens. For instance, composed image-text data is transformed into an interleaved sequence of image and text tokens, enabling the model to process them with a single backbone.

Our Lenses model builds on top of BGE-VL-MLLM [49], which itself is built on top of LLaVA-1.6 Mistral 7B [11]. In this architecture, a ViT-based visual encoder produces image tokens that are projected into the LLM space and concatenated with text tokens, so the entire multimodal input is processed by one model. For each training pair, we represent it with a *set* of embeddings rather than a single embedding by introducing special tokens that ask the LLM to emit multiple slots for a single input $\mathcal{I}$ or caption $\mathcal{S}(\mathcal{I})$. The resulting sets are meant to encode heterogeneous semantics that appear in the input data.

We follow a typical multimodal query input, but instead of producing a single global embedding for each image, we introduce *prompt* tokens that tell the MLLM to emit multiple embeddings for the same input. Concretely, given an image $\mathcal{I}$ and its image-side prompts $\mathcal{P}(\mathcal{I}) = \{(p_z, c_z)\}_{z=1}^{Z_\mathcal{I}}$, where each p_z is a short natural-language lens prompt description and $c_z \in \mathcal{C}$ is its lens label, we format the input as

$$\langle\text{instruct}\rangle\ \{\text{task\_inst}\}\ \langle\text{image}\rangle\ \{q_z^{\mathcal{V}}\}_{z=1}^{Z_\mathcal{I}}\ \texttt{[EOS]}, \quad (1)$$

where each

$$q_z^{\mathcal{V}} = p_z\ \texttt{<PROMPT>}$$

is a lens-specific query tied to category c_z. Here <PROMPT> is a learned special token in the tokenizer; we later take its final hidden state as the visual slot embedding for that prompt. The model processes all $Z_\mathcal{I}$ queries in a single pass. Because all lens-specific prompts are processed in one forward pass of the MLLM, the resulting set can encode heterogeneous semantics present in the image while staying aligned in the same retrieval space.

Textual set features. For textual features $(t_k, c_k) \in \mathcal{S}(\mathcal{I})$, we apply the same idea to captions. Each caption comes with a lens label $c_k \in \mathcal{C}$, but during encoding we expose the full lens inventory to the MLLM. For a text-only input we format

$$\langle\text{instruct}\rangle\ \{\text{task\_inst}\}\ q_t\ \texttt{[EOS]}, \quad (2)$$

where

$$q_t := t_k\ \texttt{<}\square\texttt{>}\ \texttt{<}\square\texttt{>}\ \texttt{<}\square\texttt{>}\ \texttt{<}\square\texttt{>}\ \texttt{<}\square\texttt{>}. \quad (3)$$

The final hidden states at these five category tokens form the textual slot set, aligned to the same lens inventory as the visual side. We then define a binary mask $m^{\mathcal{T}}(t_k) \in \{0,1\}^{Z_{t_k}}$

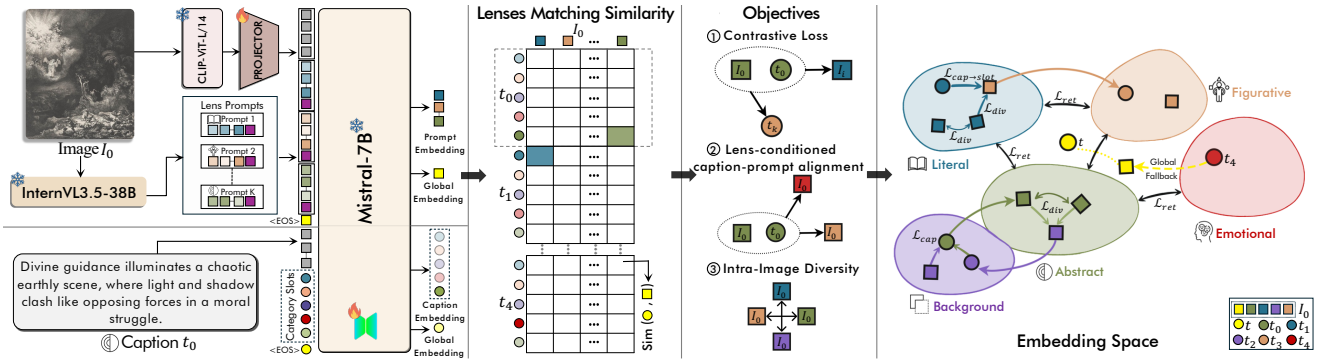

Figure 4. Overview of the Lenses model. Given an image I_0 and caption t_0, InternVL3.5-38B first imagines a set of lens prompts. The image and its prompts are fed to Mistral-7B to produce global and lens-specific image slots, while the caption is encoded under the five lenses to produce matching text slots. A lens-aware similarity module compares slots within each lens, masking cross-lens pairs and falling back to global similarity when no valid same-lens match occurs. We train with contrastive retrieval, lens-conditioned caption-prompt alignment, and intra-image diversity losses so the embedding space forms structured but interacting regions for each interpretive lens.

(with $Z_{t_k} = |\mathcal{C}|$) so that only the slot whose category matches the caption label c_k is considered active and participates in similarity computations; the remaining four slots are masked out in all losses. This lets the MLLM see the full lens vocabulary as context while ensuring that supervision flows only through the caption-annotated lens.

Slot extraction. Let $H \in \mathbb{R}^{L\times d}$ be the final hidden states of the MLLM for a sequence of length L and hidden size d. Because <PROMPT> and the five category tags <📖>, <🚶>, <🌀>, <🖼>, <🧠> are in the tokenizer, the model outputs hidden states at exactly those positions. We gather them into two padded sets:

$$E^{\mathcal{V}}(\mathcal{I}) \in \mathbb{R}^{Z_{\mathcal{I}}\times d}, \quad m^{\mathcal{V}}(\mathcal{I}) \in \{0,1\}^{Z_{\mathcal{I}}},$$

for image-side slots, and

$$E^{\mathcal{T}}(t_k) \in \mathbb{R}^{Z_{t_k}\times d}, \quad m^{\mathcal{T}}(t_k) \in \{0,1\}^{Z_{t_k}},$$

for text-side slots, where $Z_{\mathcal{I}}$ is the number of visual prompts for $|\mathcal{P}(\mathcal{I})|$ and $Z_{t_k} = |\mathcal{C}|$ is the number of textual category slots. The masks $m^{\mathcal{V}}$ and $m^{\mathcal{T}}$ indicate which rows correspond to real, active slots and which are padding or inactive. In parallel, we read a global embedding from the final token:

$$g^{\mathcal{V}}(\mathcal{I}),\ g^{\mathcal{T}}(t_k) \in \mathbb{R}^d.$$

Thus every sample exposes a dual-encoder–style global vector and a set of lens-aligned slot embeddings.

3.3. Lens matching similarity

Every image $\mathcal{I}$ and caption t_k in Lenses produces a set of lens-specific slot embeddings, $E^{\mathcal{V}}(\mathcal{I}) = \{\mathbf{v}_i\}_{i=1}^{Z_{\mathcal{I}}}$ and $E^{\mathcal{T}}(t_k) = \{\mathbf{u}_j\}_{j=1}^{Z_{t_k}}$, so a single global cosine is not expressive enough. We require a similarity that compares two sets and respects that each slot is tagged with a lens from $\mathcal{C} = \{📖, 🚶, 🌀, 🖼, 🧠\}$.

Classical set-to-set similarities give two limiting behaviors. Multiple-instance learning (MIL), as in PVSE [36], keeps only the best pair $s_{\text{MIL}}(\mathbf{S}_1, \mathbf{S}_2) = \max_{x\in\mathbf{S}_1, y\in\mathbf{S}_2}\langle x, y\rangle$, which is simple but leaves all other slots unsupervised. Fully averaging over all pairs, as in distributional approaches like PCME [5], gives dense supervision but can push all slots to overlap, leading to set collapse. SetDiv [18] lets every element pull on every other element, again averaging across lenses so that representations drift toward a single mode, and MaxMatch [2] avoids collapse via explicit matching at higher computational cost.

We introduce a lens-aware similarity that forbids cross-lens matches: a 🚶 prompt on the image side can only match a 🚶 slot on the text side, a 🖼 prompt on the image side can only match a 🖼 slot on the text side, and so on. For an image $\mathcal{I}_b$ and a caption t_n, we only allow slots that are present and from the same lens to interact. We denote

$$V_b = E^{\mathcal{V}}(\mathcal{I}_b) \in \mathbb{R}^{Z_{\mathcal{I}_b}\times d}, \qquad T_n = E^{\mathcal{T}}(t_n) \in \mathbb{R}^{Z_{t_n}\times d},$$

and let $\ell^{\mathcal{V}}_{b,i} \in \mathcal{C}$, $\ell^{\mathcal{T}}_{n,j} \in \mathcal{C}$ be the lens labels for those slots. We first form the slot-slot cosine matrix

$$S_{b,n} = V_b\, T_n^{\top} \in \mathbb{R}^{Z_{\mathcal{I}_b}\times Z_{t_n}},$$

where all rows of V_b and T_n are normalized so that entries of $S_{b,n}$ are cosines. We then keep only entries that are real slots and have the same lens. Concretely, we define a mask

$$B_{b,n}(i,j) = \begin{cases} 1, & \text{if } m^{\mathcal{V}}_{b,i} = 1,\ m^{\mathcal{T}}_{n,j} = 1 \text{ and } \ell^{\mathcal{V}}_{b,i} = \ell^{\mathcal{T}}_{n,j}, \\ 0, & \text{otherwise,} \end{cases}$$

and set $S_{b,n}(i,j) = -\infty$ whenever $B_{b,n}(i,j) = 0$. Because $m^{\mathcal{T}}_{n,j} = 1$ only for the category token that matches the caption label c_k, the remaining four category slots never participate in $s(\mathcal{I}_b, t_n)$, the contrastive losses, or the caption-slot objective. They can influence the caption representation

only as contextual tokens inside the MLLM, but are never directly optimized to match any visual slot, which prevents off-lens leakage at the similarity level. After masking, we calculate a smooth-Chamfer similarity[18]:

$$s(\mathcal{I}_b, t_n) = \frac{1}{2\alpha\, Z_{\mathcal{I}_b}} \sum_{i=1}^{Z_{\mathcal{I}_b}} \log\left(\sum_{j=1}^{Z_{t_n}} e^{\alpha S_{b,n}(i,j)}\right) + \frac{1}{2\alpha\, Z_{t_n}} \sum_{j=1}^{Z_{t_n}} \log\left(\sum_{i=1}^{Z_{\mathcal{I}_b}} e^{\alpha S_{b,n}(i,j)}\right), \tag{4}$$

where $\alpha > 0$ controls the smoothness of the max.

Some image-caption pairs in a batch have no overlapping active lenses, so all entries of $B_{b,n}$ are zero and all entries of $S_{b,n}$ are effectively masked out. In that case we fall back to the global embeddings:

$$\widetilde{S}_{b,n} = \begin{cases} s(\mathcal{I}_b, t_n), & \text{if } \exists\, i, j \;\; B_{b,n}(i,j) = 1, \\ \langle g^{\mathcal{V}}(\mathcal{I}_b),\, g^{\mathcal{T}}(t_n)\rangle, & \text{otherwise.} \end{cases}$$

Thus the lens-aware score is used whenever at least one same-lens slot is available, and the global similarity acts as a fallback when the model cannot identify valid lens overlap.

3.4. Training objective

Lenses is trained with three complementary objectives: a contrastive loss that pulls together lens-aligned image-text pairs, a slot-level alignment term that assigns each caption to the correct visual lens, and a diversity regularizer that prevents visual slots from collapsing.

Multimodal contrastive learning. We apply a standard InfoNCE loss [30] using the lens-aware similarity $\widetilde{S}_{b,n}$. For image-to-text retrieval,

$$\mathcal{L}_{\text{i}\to\text{t}} = -\frac{1}{|\mathcal{B}_{\text{img}}|} \sum_{b\in\mathcal{B}_{\text{img}}} \frac{1}{|P_b|} \sum_{n\in P_b} \log \frac{\exp(\widetilde{S}_{b,n}/\tau)}{\sum_{m=1}^{B} \exp(\widetilde{S}_{b,m}/\tau)}, \tag{5}$$

where P_b is the set of captions paired with image $\mathcal{I}_b$, $\mathcal{B}_{\text{img}}$ is the set of image indices in the batch, and τ is a temperature. Symmetrically, for text-to-image retrieval we have

$$\mathcal{L}_{\text{t}\to\text{i}} = -\frac{1}{|\mathcal{B}_{\text{txt}}|} \sum_{n\in\mathcal{B}_{\text{txt}}} \frac{1}{|Q_n|} \sum_{b\in Q_n} \log \frac{\exp(\widetilde{S}_{b,n}/\tau)}{\sum_{k=1}^{B} \exp(\widetilde{S}_{k,n}/\tau)}, \tag{6}$$

where Q_n is the set of images that match caption t_n. The retrieval term is

$$\mathcal{L}_{\text{ret}} = \mathcal{L}_{\text{i}\to\text{t}} + \mathcal{L}_{\text{t}\to\text{i}}. \tag{7}$$

Lens-conditioned caption-prompt alignment. The contrastive loss operates on $\widetilde{S}_{b,n}$ and only enforces that correct image-caption pairs are closer than mismatched ones. However, when an image exposes multiple lens-specific slots, this supervision is underconstrained: a figurative caption could be explained through the literal slot of the same image without penalty. To preserve the one-image-many-lenses factorization, we add a slot-level alignment term that operates on cosine similarity between slots,

$$C_{b,n}(i,j) = \langle \mathbf{v}_{b,i}, \mathbf{u}_{n,j} \rangle.$$

Let $\mathcal{S}_+$ denote the set of pairs (b,n) for which there exists at least one same-lens visual slot, that is, $\exists\, i, j$ such that $B_{b,n}(i,j) = 1$. For each such pair, we minimize

$$\mathcal{L}_{\text{cap}\to\text{slot}} = -\frac{1}{|\mathcal{S}_+|} \sum_{(b,n)\in\mathcal{S}_+} \frac{1}{|P_{b,n}|} \sum_{i\in P_{b,n}} \log \frac{\exp(C_{b,n}(i,j^\star)/\tau_s)}{\sum_{i'\in A_{b,n}} \exp(C_{b,n}(i',j^\star)/\tau_s)} \tag{8}$$

where $j^\star$ indexes the active textual slot (the category token whose lens equals the caption label c_k), $P_{b,n} = \{i : B_{b,n}(i,j^\star) = 1\}$ collects same-lens visual slots, $A_{b,n} = \{i : m^{\mathcal{V}}_{b,i} = 1\}$ is the set of active visual slots for $\mathcal{I}_b$, and τ_s is a slot-level temperature. This forces each caption to prefer visual slots that share its lens over other active slots.

Intra-image prompt diversity. The lens-conditioned alignment tells a caption which slot to use, but does not guarantee that visual slots are mutually informative. In practice, some lenses appear less often, and without explicit regularization their slots can drift toward whatever slot already explains most captions, typically the literal one. To prevent collapse, we add a diversity term. For each image, we compute pairwise cosine similarities between active visual slots, $\gamma_b(i,j) = \langle \mathbf{v}_{b,i}, \mathbf{v}_{b,j} \rangle$, and apply a hinge loss:

$$\mathcal{L}_{\text{div-img}} = \frac{1}{|\Omega|} \sum_b \sum_{(i,j)\in\Omega_b} \max\big(0, \gamma_b(i,j) - \alpha\big), \tag{9}$$

where $\Omega_b = \{(i,j) : i \neq j,\, m^{\mathcal{V}}_{b,i} = 1,\, m^{\mathcal{V}}_{b,j} = 1\}$ are all valid visual slot pairs and $\alpha \in [0,1)$ is a similarity margin. This encourages different prompts of the same image to occupy distinct directions, ensuring nondegenerate slots for retrieval and alignment.

Final objective. The complete training loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ret}} + \lambda_{\text{slot}}\, \mathcal{L}_{\text{cap}\to\text{slot}} + \lambda_{\text{div}}\, \mathcal{L}_{\text{div-img}}, \tag{10}$$

where λ_{slot} and λ_{div} balance cross-modal retrieval, caption-to-slot assignment, and intra-image slot diversity.

4. Experiment

4.1. Implementation details

We fine-tune our model on Lenses dataset starting from BGE-VL LLaVA-Next-1.6 Mistral-7B for retrieval. We extend the tokenizer with a prompt token `<PROMPT>` and five

	Literal				Figurative				Abstract				Background				Emotional				All			
	I→T		T→I		I→T		T→I		I→T		T→I		I→T		T→I		I→T		T→I		I→T		T→I	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Zero-shot																								
BLIP-2 [21]	40.48	65.13	37.27	61.68	13.80	31.04	10.27	26.07	9.65	22.79	5.97	16.57	11.13	24.80	8.56	20.58	12.63	27.12	9.32	21.51	6.31	17.85	3.31	13.80
CLIP [33]	46.25	66.11	39.88	64.81	18.73	37.18	14.60	34.44	13.11	26.33	8.52	23.32	7.77	19.84	8.72	20.27	13.52	28.17	10.98	25.06	8.88	20.97	3.85	16.16
CoCa [46]	58.22	76.00	49.69	73.86	21.34	42.29	14.36	34.01	13.76	30.03	8.75	21.28	13.53	27.99	8.67	20.77	17.51	34.03	11.18	24.52	11.19	25.61	4.26	18.13
OpenCLIP [12]	58.56	76.37	48.92	73.53	21.11	41.12	12.84	30.58	13.39	28.44	6.71	17.70	13.56	27.37	7.83	20.17	16.74	32.83	9.69	22.32	11.10	25.53	4.01	16.56
Jina-v4 [10]	55.33	72.74	47.56	72.42	19.87	39.93	15.75	35.92	12.53	26.72	8.59	22.39	15.29	30.83	9.10	21.58	16.88	33.54	10.54	24.79	10.93	22.72	3.81	18.35
BGE-VL-Large	49.70	69.83	53.47	78.13	14.99	31.90	16.39	38.37	9.45	22.51	9.03	23.65	13.56	28.91	13.59	30.32	13.90	29.28	13.85	30.58	9.24	20.57	4.72	20.27
BGE-VL-MLLM	63.27	81.89	72.28	85.34	28.17	48.10	36.54	58.55	17.44	33.42	22.92	41.73	7.90	17.17	23.15	42.32	14.75	28.53	26.82	47.60	70.67	88.92	36.68	55.53
Fine-Tune																								
BGE-VL-MLLM	67.34	87.73	77.51	91.42	47.52	71.46	55.52	80.59	33.52	57.35	43.14	69.56	28.61	50.93	44.98	70.25	33.80	58.16	46.60	72.42	80.89	96.16	53.88	77.84
Lenses	**80.32**	**93.60**	**81.90**	**93.60**	**56.02**	**77.64**	**62.69**	**85.15**	**41.25**	**64.34**	**51.66**	**76.59**	**33.97**	**56.54**	**48.65**	**74.67**	**40.35**	**62.95**	**51.81**	**76.84**	**89.09**	**98.18**	**58.87**	**80.74**

Table 2. Cross-modal retrieval on the Lenses test set decomposed by interpretive lens. We report Recall@1 and Recall@5 (R@K, %) for image-to-text (I → T) and text-to-image (T → I) retrieval across the literal, figurative, abstract, background, and emotional lenses; All evaluates over all captions regardless of lens. Zero-shot VLMs (top) perform well mainly on literal captions but struggle on non-literal lenses, while fine-tuning on Lenses with our lens-aware objectives (bottom) yields large gains, especially for figurative, emotional, and abstract queries and in the All setting. Best results are shown in **bold**.

lens tokens <📖>, <✨>, <🌀>, <🖼>, <🧠>. During encoding, the hidden state at <PROMPT> (and at lens tokens when present) defines slot embeddings, while the final token is used as the global embedding. At inference time, users provide only a text query; no lens label is required. The query is encoded into all five lens slots, and retrieval proceeds via lens-masked similarity with global fallback. Full training details are in the suppl.

4.2. Metrics for cross-modal retrieval

Despite long-standing awareness that image-text benchmarks often admit multiple valid matches per query, most cross-modal retrieval work still reports simple Recall@k (R@k) scores, which implicitly assume a single correct item per query [5]. Under standard R@1, a model receives full credit as soon as *any* positive caption is ranked first, even if other valid captions are pushed far down the list or surrounded by irrelevant results. This makes it impossible to distinguish a model that consistently ranks *all* relevant captions highly from one that retrieves exactly one correct caption and fills the rest of the top-k with false positives.

In our setting this limitation is amplified, because every image in Lenses is annotated under multiple interpretive lenses. To directly test whether our model learns *lens-specific* semantics, we introduce **Lens-Specific Slot Retrieval**. For a given lens c, we extract the image slot corresponding to c and match it only against text representations whose c-slot is active. We then compute image-to-text and text-to-image R@k for each of the five lenses. This slot-to-slot evaluation reveals whether, for example, the figurative slot genuinely captures figurative content, or whether all slots have collapsed to a shared representation.

Lens-aware retrieval on Lenses. Table 2 reports Recall@1 and Recall@5 for image-to-text and text-to-image retrieval on Lenses, decomposed by interpretive lens.

(a) Loss ablation

	$\mathcal{L}_{\text{ret}}$	$\mathcal{L}_{\text{cap}\to\text{slot}}$	$\mathcal{L}_{\text{div-img}}$	RSUM
1	✓			502.7
2	✓	✓		510.6
3	✓	✓	✓	513.3

(b) Text-only retrieval

K	P→C	C→P
R@1	34.13	32.73
R@5	52.15	50.31
R@10	59.83	57.82

Table 3. (a) Loss ablation: we progressively add $\mathcal{L}_{\text{cap}\to\text{slot}}$ and $\mathcal{L}_{\text{div-img}}$ on top of $\mathcal{L}_{\text{ret}}$ and report RSUM for All. (b) Text-only prompt–caption retrieval: R@K (%) for prompt-to-caption (P→C) and caption-to-prompt (C→P) matching without images.

Zero-shot vision-language models already perform reasonably well on literal captions, but their performance drops sharply for figurative, abstract, emotional, and background queries. Among them, BGE-VL-MLLM is the strongest zero-shot model, yet it still underperforms on non-literal lenses, for example achieving less than 30 R@1 on Figurative and Emotional I→T. Simply fine-tuning BGE-VL-MLLM on Lenses yields consistent gains across all lenses, especially for non-literal captions, and boosts overall All I→T R@1 from 70.67 to 80.89. Our full Lenses models provide a further jump in performance. The best literal-focused variant reaches 80.32 R@1 on Literal I→T, while our best all lens model attains 89.09 R@1 on All I→T and up to 58.87 R@1 on All T→I. Gains are largest for non-literal lenses: Figurative I→T R@1 roughly doubles compared to the best zero-shot baseline, and Emotional and Background retrieval improve by over 15 absolute points. These results show that explicitly modeling lens-specific slots and using lens-aware similarity is crucial for robust polysemous retrieval, improving non-literal understanding without sacrificing literal alignment.

ArtEmis retrieval. Table 5 reports cross-modal retrieval on ArtEmis (Human Caption). Zero-shot VLMs achieve very low recall on this affective benchmark, where the best R@1 is 5.52 for I→T and 4.84 for T→I with SigLIP-2, indicating that standard training under-represents emotional

Method	LC@10	All@10	Lens DCG@10	Caption DCG@10
BGE-VL-MLLM	48.2	8.2	50.5	45.0
BGE-VL-MLLM (Fine-Tune)	66.5	21.4	65.8	60.0
Lenses	**76.8**	**36.8**	**75.6**	**70.2**

Table 4. Lens-aware retrieval metrics on the multi-lens setup (5500 images). LensCoverage@10 (LC@10) measures the fraction of an image's lenses retrieved in the top-10; AllLenses@10 (All@10) is the percentage of images for which *all* annotated lenses appear in the top-10; lens DCG@10 is a lens-level DCG that gives credit only to the first hit per lens. Best results per column are shown in **bold** and second-best results are underlined.

and non-literal content. Fine-tuning BGE-VL-MLLM on Lenses substantially boosts performance, raising I→T R@1 to 14.03 and T→I R@1 to 12.05. Adding lens-aware training with Lenses boosts the performance even more: the All-lens model reaches 15.67/13.44 R@1 (I→T/T→I), while the Literal-focused variant attains the best overall scores with 17.32/13.51 R@1. The Emotional-only variant is competitive, especially on T→I, and all Lenses variants improve R@5 and R@10 as well. Overall, our lens-conditioned supervision on Lenses dataset leads to representations that better capture the affective and non-literal semantics.

Lens-aware retrieval and why images still matter. Table 3 evaluates our loss design and text-only prompt-caption matching. The loss ablation (Tab. 3a) shows that adding $\mathcal{L}_{\text{cap}\to\text{slot}}$ and $\mathcal{L}_{\text{div-img}}$ on top of $\mathcal{L}_{\text{ret}}$ steadily increases RSUM from 502.7 to 513.3. The text-only retrieval (Tab. 3b) shows that the learned lens-aware space supports non-literal retrieval without images (34.13 R@1 for P→C and 32.73 R@1 for C→P), but these scores are far below our image-text retrieval on Lenses (Tab. 2), where All R@1 exceeds 80. Table 4 further shows that, even after fine-tuning, BGE-VL-MLLM reaches only 66.5 LC@10 and 21.4 All@10, while Lenses attains 76.8 and 36.8 and also improves lens and caption DCG@10. Thus, prompts alone can retrieve related captions, but joint training with images is what ties slots to grounded lenses that consistently cover an image's diverse perspectives.

Does performance just come from a bigger InternVL? Table 6 tests whether our gains could be explained by the size of the InternVL3.5 backbone used to generate prompts. We keep our retrieval architecture and training fixed and swap InternVL3.5-4B, 8B, and 14B. Scaling from 4B to 14B changes macro R@K only marginally (around 0.5–1 point in R@1 for both I→T and T→I), suggesting limited benefit from backbone capacity alone. In contrast, the InternVL3.5-38B variant, which pairs the same retrieval head with lens prompts and slot-masked similarity, improves R@1 to 48.30 (I→T) and 45.40 (T→I), a 3–4 point gain over the best unprompted setting. Since all rows use the same retrieval head and training recipe, this indicates that our improvements are not simply an artifact of a larger InternVL model, but stem from lens-aware prompting and our slot-based objective.

Model	I → T			T → I		
	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot						
BLIP-2 [22]	1.62	5.23	7.86	2.46	9.65	13.57
BGE-VL-Large	2.25	6.67	9.74	3.86	14.24	19.62
CLIP [33]	3.83	9.59	13.04	3.76	13.72	19.07
CoCa [46]	4.35	10.15	13.67	4.57	16.25	22.04
Jina-v4 [10]	2.74	7.62	10.93	2.18	9.04	13.07
Onepiece [42]	5.34	13.44	18.08	4.78	17.75	24.17
OpenCLIP [12]	3.86	9.58	13.04	3.53	13.59	19.14
SigLIP-2 [39]	5.52	12.45	16.53	4.84	17.74	23.92
Fine-tuned on Lenses Dataset						
BGE-VL-MLLM	14.03	28.87	36.81	12.05	25.53	33.14
Lenses - All	15.67	31.85	41.34	13.44	26.60	37.98
Lenses - 🕮	**17.32**	**33.95**	**42.71**	**13.51**	**29.81**	**40.34**
Lenses - 🧠	16.28	31.91	39.84	12.67	28.45	39.56

Table 5. Cross-modal retrieval results on ArtEmis (Human Caption). We report Recall@K (R@K, %) for image to text (I→T) and text to image (T→I) retrieval. Zero shot baselines are shown in the top block, while the bottom block fine-tunes BGE-VL-MLLM on the Lenses dataset, either on all lenses or on the Literal 🕮 and Emotional 🧠 subsets. Best results per column are shown in **bold** and second-best results are underlined.

Model	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
InternVL3.5-4B	45.44	66.04	73.53	40.99	63.03	71.27
InternVL3.5-8B	45.04	65.19	73.00	40.59	62.86	71.08
InternVL3.5-14B	44.77	65.12	73.12	41.58	63.69	71.88
InternVL3.5-38B	**48.30**	**68.04**	**75.46**	**45.40**	**66.41**	**73.88**

Table 6. Ablation of InternVL3.5 model size under our model. We report macro-averaged Recall@K (R@K, %) over the five lenses using slot-masked similarity with global fallback. Scaling the captioning backbone from 4B to 14B yields only marginal and non-monotonic changes in retrieval, while the 38B variant with lens prompts and slot-masked similarity achieves the best performance. Full table in the supplementary.

5. Conclusion

We presented Lenses, a dataset, model, and evaluation framework that treat image-text retrieval as matching across multiple interpretive lenses rather than a single literal meaning. By annotating each image with captions and prompts for five lenses and training a multi-prompt dual encoder with a lens-aware similarity module and diversity regularization, our approach learns structured, lens-specific embeddings while avoiding slot collapse. New lens-aware metrics, including lens-specific slot retrieval, reveal performance differences that standard Recall@k obscures and show that our model better recovers both literal and non-literal readings than strong baselines. We hope Lenses provides a foundation for future work on polysemous vision-language understanding, richer lens taxonomies, and applications that must adapt an image's reading to user goals and context.

6. Acknowledgements

We acknowledge Advanced Research Computing at Virginia Tech for providing computational resources and technical support that have contributed to the results reported in this paper. We also acknowledge support from the Commonwealth Cyber Initiative Southwest Virginia Node for professional development and travel assistance that helped make participation in this work possible. We thank the reviewers for their valuable comments, which helped improve the paper.

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579, 2021. 1

[2] Hani Alomari, Anushka Sivakumar, Andrew Zhang, and Chris Thomas. Maximal matching matters: Preventing representation collapse for robust cross-modal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31769–31785, 2025. 2, 3, 5

[3] David Chan, Austin Myers, Sudheendra Vijayanarasimhan, David Ross, and John Canny. Ic3: Image captioning by committee consensus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8975–9003, 2023. 1

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2

[5] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8415–8424, 2021. 2, 5, 7

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

[7] Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[8] Rachel Giora. Literal vs. figurative language: Different or equal? *Journal of pragmatics*, 34(4):487–506, 2002. 1

[9] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7181–7189, 2018. 2

[10] Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, et al. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550, 2025. 2, 7, 8

[11] Hugging Face. LLaVA-v1.6-Mistral-7b-hf. https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf, 2024. 4

[12] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021. 7, 8

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[14] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 4

[15] 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. *arXiv preprint arXiv:2407.12580*, 2024. 2

[16] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020. 2

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2

[18] Dongwon Kim, Namyup Kim, and Suha Kwak. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23422–23431, 2023. 2, 3, 5, 6

[19] Jinhyuk Lee, Zhuyun Dai, Sai Meher Karthik Duddu, Tao Lei, Iftekhar Naim, Ming-Wei Chang, and Vincent Zhao. Rethinking the role of token retrieval in multi-vector retrieval. *Advances in Neural Information Processing Systems*, 36:15384–15405, 2023. 2

[20] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 2

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International*

conference on machine learning, pages 12888–12900. PMLR, 2022. 2, 7

[22] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 8

[23] Valerii Likhosherstov, Anurag Arnab, Krzysztof Marcin Choromanski, Mario Lucic, Yi Tay, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *Transactions on Machine Learning Research*, 2023. 2

[24] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 3

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[27] Saif Mohammad and Svetlana Kiritchenko. Wikiart emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018. 3

[28] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. Transfer learning approaches for building cross-language dense retrieval models. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 382–396, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[29] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 2

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[31] W. N. Org. Wikiart dataset. https://www.wikiart.org/, 2025. Accessed July 2025. 3

[32] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 3

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 7, 8

[34] Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1–23, 2025. 1

[35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3

[36] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1979–1988, 2019. 2, 3, 5

[37] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14398–14409, 2024. 2

[38] Christopher Thomas and Adriana Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 317–335. Springer, 2020. 2

[39] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 8

[40] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 3

[41] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *AAAI Conference on Artificial Intelligence*, 2018. 3

[42] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023. 8

[43] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023. 2

[44] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. Universal weighting metric learning for

cross-modal matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13005–13014, 2020. 2

[45] Ron Yosef, Yonatan Bitton, and Dafna Shahaf. Irfl: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, 2023. 1

[46] 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*, 2022. 7, 8

[47] 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*, pages 11975–11986, 2023. 2

[48] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 59403–59420. PMLR, 2024. 3

[49] 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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19076–19095, 2025. 2, 4