Lenses: Toward Polysemous Vision–Language Understanding

Supplementary Material

This supplementary document provides additional analyses and implementation details that complement the main paper on Lenses. We begin by motivating our five-lens taxonomy (Sec. 1) and describing the full training setup, optimization hyperparameters, and evaluation metrics (Sec. 7). Sec. 9 presents a human validation study of captions and lens labels. We then report zero-shot retrieval results (Sec. 2), backbone-scaling experiments (Sec. 6), and ablations on lens-masked similarity, visual prompt density, and caption-generator dependence (Secs. 3–5). Sec. 11 isolates the role of architecture versus data diversity. Finally, we detail the prompt design and validation pipeline (Sec. 12), analyze how caption lenses route to image lenses and slots (Sec. 8), and provide qualitative retrieval examples.

1. Our choice of five lenses

The lenses we use were chosen for clear reasons. Earlier works [1, 3, 13, 16] are aware that visuals can hold more than one type of meaning, but they usually study these meanings from one angle at a time, such as focusing only on literal content, or only on figurative, abstract, background aspects, or emotional. We began by adding the *literal* and *abstract* lenses to separate concrete details from broader ideas. The Artemis work [1] showed that emotional meaning needs its own space, so we added the *emotional* lens. Research on figurative and symbolic meaning, including studies on visual associations [3] and recent work on figurative understanding [13], as well as the use of external knowledge for decoding visual symbolism [16], showed the need for a *figurative* lens. We then added the *background* lens to capture context, setting, and style. With these five lenses we introduce the idea of polysemous understanding, but this does not mean the concept is limited to only five lenses. They are strong examples based on detailed prior work that show how many layers an image can carry. Our work brings these viewpoints together and applies them to multimodal retrieval so the system can use all these layers of meaning instead of just one.

2. Zero-shot retrieval benchmarks

To check that learning lens-aware, polysemous representations does not break conventional literal retrieval, we evaluate Lenses in a zero-shot setting on COCO [8] and Flickr30K [11], whose captions are mostly literal (Tables 1 and 2). On both benchmarks, the full multi-lens model Lenses All lags behind the strongest literal specialists such as VLM2Vec [7] and the BGE-VL-MLLM backbone [18], especially for T→I, which is unsurprising given that our model focuses on diverse retrieval and allocates capacity to non-literal lenses and is not tuned for these datasets. At the same time, Lenses All remains clearly competitive with CLIP-style and MagicLens baselines, showing that explicitly modeling figurative, emotional, abstract, and background content does not catastrophically harm performance in the standard literal retrieval. More importantly, when we restrict retrieval to the Literal lens, performance on I→T nearly matches or even surpasses the underlying backbone (*e.g.* 75.5 vs. 75.4 R@1 on COCO and 84.7 vs. 80.1 R@1 on Flickr30K), while maintaining strong recall at higher ranks. Together, these results tell a consistent story: Lenses is not designed to chase state-of-the-art numbers on classic literal benchmarks, but it preserves strong literal retrieval when asked to focus on that lens, while gaining the ability to operate in richer, non-literal interpretive settings that conventional models cannot target.

Model	I → T			T → I		
	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot						
OpenAI CLIP-B/16	33.1	58.4	69.0	52.4	76.8	84.7
Open CLIP-B/16	42.3	66.7	77.1	59.4	81.8	88.6
EVA-02-CLIP-B/16	42.2	66.9	76.3	58.7	80.7	88.2
OpenAI CLIP-L/14	36.5	61.0	71.1	56.3	79.3	86.7
Open CLIP-L/14	46.1	70.7	79.4	62.1	83.4	90.3
EVA-02-CLIP-L/14	47.5	71.2	79.7	63.7	84.3	90.4
MagicLens-B	48.9	73.9	82.5	64.8	85.5	91.2
MagicLens-L	53.1	77.4	84.9	67.7	87.6	92.7
VLM2Vec [7]	**75.7**	–	–	**73.1**	–	–
BGE-VL-MLLM [18]	75.4	**94.90**	**98.30**	69.91	**91.80**	**96.12**
Lenses - All	67.6	90.3	94.6	53.76	82.62	91.46
Lenses - Literal	75.5	93.6	97.4	60.1	86.3	93.3

Table 1. Zero-shot cross-modal retrieval results on COCO [8] (5K test set). We report Recall@K (R@K, %) for image to text (I→T) and text to image (T→I) retrieval. CLIP [12] baselines are taken from prior work, and MagicLens results are from our reproduced CoCa checkpoints, keeping only the best variant per model family. VLM2Vec [7], BGE-VL-MLLM [18], and our Lenses variants are evaluated in the same zero-shot setting. Best result in each column is in bold; second best is underlined.

3. Are our gains from more embeddings or better structure?

To isolate the source of our performance gains, we must determine if they stem merely from increased representational capacity (*i.e.* using multiple embeddings per image) or from the specific structural constraints of our lens-aware masking.

Model	I → T			T → I		
	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot						
OpenAI CLIP-B/16	62.1	85.6	91.8	81.9	96.2	98.8
Open CLIP-B/16	69.8	90.4	94.6	86.3	97.9	99.4
EVA-02-CLIP-B/16	71.2	91.0	94.7	85.7	96.7	98.9
OpenAI CLIP-L/14	65.2	87.3	92.0	85.2	97.3	99.0
Open CLIP-L/14	75.0	92.5	95.6	88.7	98.4	99.2
EVA-02-CLIP-L/14	77.3	93.6	96.8	89.7	98.6	99.2
MagicLens-B	76.2	93.7	96.5	87.9	97.7	99.5
MagicLens-L	79.7	95.0	97.4	89.6	98.7	99.4
VLM2Vec [7]	80.3	95.0	97.4	**94.6**	**99.5**	**99.8**
BGE-VL-MLLM [18]	80.1	95.9	**98.6**	78.2	93.3	96.6
Lenses - All	77.8	94.2	96.3	63.42	86.0	92.1
Lenses - 🕮	**84.7**	**96.0**	98.0	69.3	89.8	94.4

Table 2. Zero-shot cross-modal retrieval results on Flickr30K. We report Recall@K (R@K, %) for image to text (I→T) and text to image (T→I) retrieval. As a general multimodal representation model, VLM2Vec achieves strong T→I and I→T scores compared to existing CLIP-like models. Baseline numbers, including CLIP, MagicLens [17], and VLM2Vec, are sourced from [7]. Best result in each column is in bold, second best is underlined.

In Table 3, we perform an ablation where the underlying image and text embeddings are kept fixed, and only the similarity aggregation mechanism is varied. We compare three paradigms: **global baselines**, which mean doing a standard retrieval using a single pooled embedding per modality (with and without image-side prompts), **unstructured set similarity**, a "No-mask" variant that computes smooth Chamfer similarity across *all* available slots, ignoring lens labels, and **structured lens-masking similarity**, our proposed method, which restricts matching to lens-consistent slots, with and without a global fallback.

Capacity vs. Structure. First, we observe that all multi-slot variants significantly outperform the Global-only baselines (e.g., jumping from 55.5 to 58.4 R@1 in T→I). This confirms that compressing complex multimodal semantics into a single vector creates a bottleneck, and that set-based representations are inherently superior.

Crucially, however, structure matters. The *Lens-masked* variant outperforms the *No-mask* variant (87.8 vs 87.1 I→T R@1), indicating that enforcing semantic consistency between slots helps filter out noisy, irrelevant matches. The strongest performance is achieved by the full model (*Lens-masked + Global*), which combines the precision of lens-specific matching with the robustness of a global safety net.

Fallback Analysis. Finally, we analyze the behavior of the fallback router in Table 4. The fallback mechanism is triggered when no compatible lens slots overlap between query and target. We find this occurs most frequently for the 🕮 Literal lens (13.4%), likely because literal queries can be broad and may not always align with a specific slot configuration. Conversely, non-literal lenses rely on fallback rarely (3.1% for 🧠 Emotional), suggesting that our specialized slots successfully capture the majority of the signal for abstract and affective concepts. This confirms that the fallback path acts strictly as a safety net for edge cases, rather than a primary driver of performance.

Similarity Variant	Image → Text		Text → Image	
	R@1	R@5	R@1	R@5
Global Baseline				
Global only (w/o prompts)	81.4	96.2	53.2	78.2
Global only (w/ prompts)	84.1	96.5	55.5	80.5
Structured Similarity				
No-mask smooth Chamfer	87.1	97.8	58.4	80.7
Lens-masked smooth Chamfer	87.8	98.0	59.1	81.2
Full Model (Lens-masked + Global)	**88.3**	**98.2**	**59.4**	**81.9**

Table 3. **Disentangling Capacity vs. Structure.** We ablate the similarity module while keeping embeddings fixed. *No-mask* computes similarity across all slots (high capacity, low structure). *Lens-masked* restricts matching to same-lens slots (high structure). The results show that lens alignment adds value beyond simply increasing the number of embeddings.

Lens Category	Fallback Rate (%)
🕮 Literal	13.4
Figurative	10.3
Abstract	7.1
Background	4.8
🧠 Emotional	3.1

Table 4. **Global Fallback Usage.** Percentage of queries in the Lenses test set where the model routes to the global embedding because no lens-consistent slots were found. Low usage on non-literal lenses indicates high coverage by our specialized slots.

4. How dense do visual prompts need to be?

In this experiment, we decouple the contribution of the learned lens embeddings from the explicit visual prompts. Specifically, we investigate whether the model successfully learns a "lens direction" purely through the tokenizer or if it requires the semantic grounding provided by image-conditioned prompts. Table 5 presents the performance as we vary z, the number of visual prompts injected per lens.

1. **Token-only ($z = 0$):** Relying solely on learned lens tokens yields respectable performance on the 🕮 Literal lens, suggesting the base model already possesses strong literal grounding. However, performance on non-literal lenses suffers significantly (e.g., Figurative I→T R@1 drops to 39.8). This indicates that while learned tokens can signal *intent*, they lack the semantic resolution to bridge the gap to complex abstract concepts on their own.

Config (z)	Literal				Figurative				Abstract				Background				Emotional				All-Lenses			
	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
z=0 (Token-Only)	67.4	86.7	76.2	90.2	39.8	63.6	51.2	74.8	28.0	49.4	38.9	64.6	24.5	46.1	39.6	64.5	28.3	50.2	41.8	66.8	77.3	94.3	50.0	72.5
z=1 (Top-1 Prompt)	**82.0**	**93.8**	81.1	92.4	54.5	74.7	60.8	82.0	37.8	59.8	46.8	71.5	31.8	53.1	45.6	69.6	37.3	59.8	49.6	72.6	88.2	97.7	57.3	78.1
z>1 (All-Prompts)	80.3	93.6	**81.9**	**93.6**	**56.0**	**77.6**	**62.7**	**85.2**	**41.3**	**64.3**	**51.7**	**76.6**	**34.0**	**56.5**	**48.7**	**74.7**	**40.4**	**63.0**	**51.8**	**76.8**	**89.1**	**98.2**	**58.9**	**80.7**

Table 5. **Impact of Visual Prompt Density.** We evaluate retrieval performance on the Lenses test set as we vary z, the number of visual prompts provided per lens. While the model can perform basic literal retrieval with learned tokens alone (z=0), explicit visual prompts ($z \geq 1$) are critical for unlocking performance on abstract and figurative lenses. **Bold**: Best; Underline: Second Best.

2. **Single Prompt ($z = 1$):** Injecting just a single visual prompt per lens catalyzes a dramatic performance leap. Figurative retrieval jumps by nearly 15 absolute percentage points (39.8 $\rightarrow$ 54.5), and the overall All-Lenses retrieval improves from 77.3 to 88.2. This confirms that visual prompts serve as essential semantic anchors that guide the model's attention to specific visual attributes.
3. **Full Density ($z > 1$):** Utilizing all available prompts yields further gains. While we observe diminishing returns compared to the initial jump from $z = 0$ to $z = 1$, the improvement remains robust for the most subjective lenses (*e.g.* Emotional and Abstract). This suggests that for highly polysemous concepts, a single description is insufficient; a "cloud" of diverse prompts is necessary to fully span the semantic manifold of the lens.

5. Are our gains just an artifact of InternVL captions?

Table 6 directly targets the concern that our improvements might simply reflect circularity in the data pipeline, since both the training captions and the image-side prompts are generated by InternVL3.5. To stress-test this, we freeze the Lenses retriever model trained on InternVL3.5-38B captions and prompts, keep the InternVL3.5-38B image-side prompts fixed at test time, and vary only the captioning MLLM used to form the text queries and candidates, using captions from Qwen3VL-30B-A3B [15], Grok 4.1 Fast [14], or InternVL3.5-38B [2]. Across all three caption generators, Lenses consistently outperforms the corresponding fine-tuned BGE-VL-MLLM baseline in the All setting and remains competitive or better on the more challenging non-literal lenses (Figurative, Abstract, Emotional), even though it has only ever seen InternVL3.5-38B captions during training. Moreover, All R@1/R@5 with Qwen3VL-30B-A3B and Grok 4.1 Fast captions is comparable to, and in several cases slightly higher than, the InternVL caption condition, indicating that our model does not rely on InternVL-specific phrasing. This robustness to swapping the caption generator, together with the consistent margins over the fine-tuned BGE-VL-MLLM baseline, suggests that Lenses has learned a genuinely useful lens-aware representation that transfers across independently generated captions rather than overfitting to stylistic quirks of a single MLLM, substantially mitigating circularity concerns.

6. Impact of Backbone Scaling

A natural concern is that our ability to retrieve non-literal semantics might simply reflect the strength of the captioning backbone used to generate prompts, rather than the structure of the retriever itself. Perhaps any sufficiently large InternVL-style model, plugged in as a prompt generator at test time, would make a fixed retriever look good on Lenses. To disentangle captioning capacity from retrieval supervision, we run a backbone-scaling study where the Lenses retrieval architecture is completely frozen and only the multimodal backbone that produces prompts and lens labels is changed.

In Table 7, we take the Lenses retriever trained once on the Lenses dataset (whose captions and image-side prompts were generated with InternVL3.5-38B) and re-evaluate it while varying the model that supplies the lens-conditioned prompts and lens labels at inference time. For each backbone we issue exactly the same instruction template; we then feed the resulting prompt and category lens label into the frozen retriever. The prompt text is appended to the image, and the predicted lens determines which slots are allowed to match under our lens-masked similarity. Thus every row in the table uses exactly the same retrieval head, parameters, and training data; the only difference is which captioning backbone (BGE-VL-MLLM or InternVL3.5 with 4B, 8B, 14B, or 38B parameters) is used to generate prompts at test time.

Scaling this captioning backbone gives only limited benefits. Moving from 4B to 14B leads to modest improvements and, in several cases, worse T→I performance than the original configuration: for Figurative, Abstract, Background, and Emotional lenses, T→I R@1 with 4B/8B prompts often trails the BGE-VL-MLLM row. The smaller InternVL variants are particularly brittle on non-literal lenses in the T→I direction, where 4B/8B/14B prompts often trail the BGE-VL-MLLM row despite the retriever being identical across all configurations.

Clear gains appear only when using the 38B captioning backbone. Prompts from InternVL3.5-38B substantially

	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
Qwen3VL-30B-A3B (Captions)																								
BGE-VL-MLLM	75.45	91.93	**94.84**	99.43	53.47	76.14	58.48	81.31	31.02	52.39	45.01	70.98	37.70	60.74	57.57	**80.70**	36.82	58.60	52.29	76.96	84.83	97.16	61.12	81.71
Lenses	**83.68**	**94.72**	94.08	**99.50**	**58.34**	**78.44**	**61.83**	**83.23**	**35.42**	**55.95**	**48.77**	**73.72**	**40.13**	**61.93**	**57.63**	80.61	**38.76**	**60.59**	**54.19**	**77.73**	**89.29**	**98.06**	**63.07**	**82.96**
Grok 4.1 Fast (Captions)																								
BGE-VL-MLLM	69.36	87.80	86.25	96.56	60.88	82.52	57.34	79.48	27.11	48.80	43.74	70.91	26.20	46.60	48.76	73.49	31.13	52.71	46.13	71.34	83.02	96.58	56.59	78.58
Lenses	**80.80**	**93.29**	**88.42**	**97.16**	**66.86**	**86.33**	**60.88**	**82.29**	**31.17**	**52.68**	**48.40**	**74.21**	**28.45**	**49.49**	**49.83**	**75.13**	**33.76**	**55.16**	**48.68**	**73.68**	**88.89**	**98.05**	**59.68**	**80.96**
InternVL3.5-38B (Captions)																								
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 6. Effect of captioning backbone on retrieval performance on the Lenses test set. For each caption generator (Qwen3VL-30B-A3B, Grok 4.1 Fast, InternVL3.5-38B), we report image-to-text (I→T) and text-to-image (T→I) R@1 and R@5 across the five lenses and in the All setting, comparing a BGE-VL-MLLM baseline fine-tuned on the Lenses training split to our full model (Lenses). Within each caption generator block, the higher score between BGE-VL-MLLM and Lenses is shown in **bold**.

improve retrieval on Abstract and Emotional captions (e.g., Abstract T→I R@1 rises from around 30–33% with 4B/8B/14B to 51.7%), and they boost Figurative T→I performance as well. However, even with 38B prompts, a pronounced gap remains between literal and non-literal retrieval: Literal T→I R@1 reaches 81.9%, whereas Figurative, Abstract, Background, and Emotional lenses remain in the 48–63% range. Because the retriever, loss, and masking are fixed across all rows, this experiment shows that simply scaling the captioning backbone used to generate prompts does not by itself solve polysemous retrieval on Lenses ; strong non-literal performance still requires explicit lens-aware supervision in the retriever rather than relying solely on a more powerful prompt generator.

7. Implementation Details

Architecture and Parameterization. We build Lenses upon `BAAI/BGE-VL-MLLM-S1` [18], a LLaVA-Next style architecture capable of interleaved vision–language processing. Specifically, we utilize a frozen Mistral 7B language backbone and a frozen CLIP-based vision tower. To adapt the model, we introduce Low-Rank Adaptation (LoRA) [6] to the language backbone (rank $r = 16$, $\alpha = 32$, dropout 0.1) and keep the image projector fully trainable. While the base model supports variable resolutions, we standardize inputs to 512×512 to maintain consistent token sequence lengths [18].

Training Protocol. We train Lenses for approximately 10,000 steps using the AdamW optimizer. Each GPU processes a mini-batch of 20 examples, and we use 8 GPUs in parallel, giving 160 samples per forward pass. We then accumulate gradients over 4 such micro-steps before each optimizer update, resulting in an effective batch size of 640 samples per optimization step ($20 \times 8 \times 4$). We employ a cosine annealing schedule, decaying the learning rate from 5×10^{-5} to 1×10^{-6} with a weight decay of 0.01. To stabilize training, we clip the global gradient norm at 1.0. Input processing handles vision tokens and image-side prompts with a maximum sequence length of 3,500 tokens, while caption-only inputs are truncated to 256 tokens. Regarding loss configuration, the multi-positive supervised-contrastive retrieval loss operates with a temperature $\tau = 0.07$. For our lens-aware components, we use a temperature of $\alpha = 16.0$ and a denominator of 2.0 for the smooth Chamfer similarity module. Based on the ablation study in Table 8, we set the loss weights for lens-conditioned alignment (λ_{slot}) and intra-image prompt diversity (λ_{div}) to 0.05 and 0.01, respectively. Experiments were conducted on a single node with 8 NVIDIA H200 GPUs (141 GB HBM), utilizing mixed precision and gradient checkpointing to optimize memory efficiency.

Inference and efficiency. At inference time, users provide only a query; no external lens label is required. The query is encoded into lens-conditioned slots, and retrieval is performed over cached image embeddings using lens-masked similarity with a global fallback when needed. This keeps retrieval in the bi-encoder setting: each image is encoded once into a small set of embeddings in a single pass, and scoring remains a dot-product-style aggregation over cached embeddings. In practice, this introduces about a $5\times$ embedding storage overhead relative to a single global embedding while preserving efficient retrieval.

Token Initialization. To explicitly model diverse interpretations, we augment the tokenizer with a generic marker, `<PROMPT>`, and five specific lens tokens (*e.g.* `<EMOTIONAL>`, `<ABSTRACT>`). To facilitate rapid convergence, we initialize these new tokens by copying the learned embedding of the end-of-sequence token (`</s>`) and adding small isotropic Gaussian noise ($\sigma = 0.1$, scaled by feature dimension) to break symmetry. To ensure consistency between input and output representations, we mirror these initialized vectors to the output embedding matrix, training them jointly with the model.

Baselines. For a fair comparison, we implement a **Fine-**

Backbone	Literal				Figurative				Abstract				Background				Emotional			
	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
BGE-VL-MLLM	67.3	87.7	77.5	91.4	47.5	71.5	55.5	80.6	33.5	57.4	43.1	69.6	28.6	50.9	45.0	70.3	33.8	58.2	46.6	72.4
InternVL3.5-4B	80.6	93.7	69.6	86.2	50.2	71.7	47.3	71.0	32.0	53.3	30.6	54.5	30.0	52.7	27.6	50.3	34.4	58.8	30.0	53.2
InternVL3.5-8B	81.1	93.8	69.3	86.1	48.9	70.5	46.3	70.0	31.2	51.9	28.5	52.7	27.2	48.9	25.1	47.4	36.8	60.9	33.7	58.1
InternVL3.5-14B	**82.0**	**94.2**	74.0	89.3	**56.8**	**77.7**	55.6	77.0	34.1	56.8	33.3	58.7	29.6	52.0	25.4	48.6	34.6	58.4	31.4	54.5
InternVL3.5-38B	80.3	93.6	**81.9**	**93.6**	56.0	77.6	**62.7**	**85.2**	**41.3**	**64.3**	**51.7**	**76.6**	**34.0**	**56.5**	**48.7**	**74.7**	**40.4**	**63.0**	**51.8**	**76.8**

Table 7. **Effect of Backbone Scaling.** We compare lens-specific retrieval performance when a single frozen Lenses retriever is paired at test time with different captioning backbones used to generate lens-conditioned prompts and labels. Scaling the captioning backbone to 38B provides absolute gains, but smaller backbones (4B–14B) often underperform the BGE baseline in Text→Image retrieval, indicating that parameter count alone does not solve polysemous mapping. The retriever is fine-tuned once on the Lenses dataset and kept fixed; only the prompt generator backbone varies across rows. **Bold**: Best; Underline: Second best.

λ_{slot}	0.1	**0.05**	0.01
RSUM	463.4	**472.3**	468.9

λ_{div}	0.1	0.05	**0.01**
RSUM	453.2	461.5	**465.6**

Table 8. **Hyperparameter sensitivity.** We report Recall Sum (RSUM) on the validation set. To select the slot weight (λ_{slot}) and diversity weight (λ_{div}), we train a scaled-down version of our model on 20% of the training data and sweep each coefficient separately. When tuning λ_{slot} we use the contrastive retrieval loss plus the slot loss with λ_{div}=0; when tuning λ_{div} we use the contrastive loss plus the diversity loss with λ_{slot}=0.

tuned BGE-VL-MLLM baseline. This model shares the identical architecture, data pipeline, and optimization hyperparameters as Lenses. However, it is trained solely with a global contrastive loss over image and text embeddings, omitting the lens-specific prompt sets and the Chamfer similarity module.

Evaluation metrics. We evaluate along two axes: instance-level retrieval accuracy and lens-specific semantic diversity.

Retrieval accuracy.

- **Standard Recall@K (R@K).** The conventional cross-modal retrieval metric. For image-to-text (I→T), a query is counted as correct if any caption associated with the ground-truth image appears in the top-K retrieved captions. For text-to-image (T→I), it is correct if the ground-truth image appears in the top-K retrieved images, regardless of which lens the caption belongs to.
- **Per-category Recall@K (Lens R@K).** A per-lens breakdown of retrieval performance with the full gallery kept intact. For each lens $\ell \in \{\text{Literal, Figurative, Abstract, Background, Emotional}\}$, we restrict the set of queries to those annotated with lens ℓ, but we do not modify the retrieval pool: all images (or captions) in the split remain as candidates, with their original positives and negatives. We then compute standard R@K on this restricted query set. This measures how well the model serves queries from each lens without artificially simplifying the candidate set.

Lens diversity. To measure how well the model recovers the full spectrum of interpretations for each image, we treat all annotated captions across lenses as positives and report:

- **LensCoverage@10 (LC@10).** For each image, we compute the fraction of its annotated lens categories that appear at least once among the top-10 retrieved positive captions, then average this fraction over images.
- **AllLenses@10 (All@10).** The percentage of images for which every annotated lens category is represented by at least one positive caption within the top-10 retrieved items.
- **Lens DCG@10.** A discounted cumulative gain defined over lens categories rather than individual captions. Each lens contributes at most once, via the highest-ranked caption belonging to that lens; additional captions from the same lens do not increase the score, explicitly rewarding diversity over redundancy.
- **Caption DCG@10.** Standard DCG computed over all positive captions, providing a holistic view of ranking quality without lens-specific diversity penalties.

8. Do caption lenses route to consistent image lenses and slots?

Figure 1 probes how our lens-aware similarity actually organizes the image-side slots. For every caption in the Lenses validation set, we record which image slot achieves the highest score and summarize the winners by caption lens. The left confusion matrix shows that captions strongly prefer the matching image lens (Literal, Emotional, and Background all peak on their own lens), while the remaining mass is concentrated on semantically adjacent lenses such as Figurative-Abstract, rather than being spread uniformly. The right panel looks inside the slots themselves: each caption lens routes to a small, characteristic subset of slots, with different lenses specializing in different indices and only partial overlap between them. Together, these patterns suggest that the model has learned a soft but structured partition of the image representation space, where different lenses claim

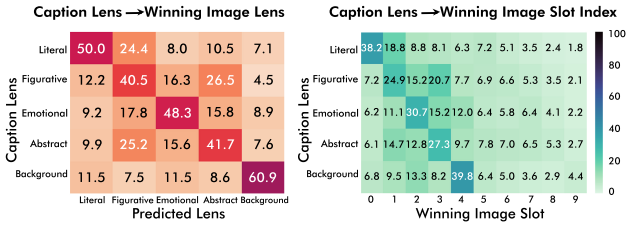

Figure 1. Lens-aware routing on the Lenses validation set. Left: captions usually select image slots from the matching lens, with mass concentrated on the diagonal and nearby related lenses. Right: each caption lens favors a small subset of image slots, indicating that slots specialize by lens instead of collapsing to a single global embedding. Values indicate the percentage of captions in each cell.

different slots rather than collapsing all semantics into one global embedding.

9. Human validation of captions and lens labels

To evaluate the quality of the automatic annotations in Lenses we conduct a human study on a subset of the test split. We draw 367 image caption pairs in a balanced way across the five lenses, with 72 Literal, 84 Figurative, 68 Emotional, 78 Abstract, and 65 Background examples. Each item is evaluated by three annotators selected from a group of five trained individuals, which gives 1101 total judgments. For every item, the annotators see the image, the caption, and the dataset lens label. They answer four questions. The first measures caption quality on a four-point scale from Bad to Excellent. The second asks whether they agree with the assigned category. The third asks them to choose the best-fitting lens from the five options or mark the item as none or unsure. The fourth asks whether they used any outside information. Before starting the task, all annotators read a one-page guide with definitions and examples for each lens and complete a short practice stage.

Figure 2 shows the distribution of caption quality scores. After removing one unclear answer, we keep 1100 valid ratings across 366 items. In total 88.0 percent of all ratings fall in the Good or Excellent range. For 90.2 percent of items, at least two annotators give a Good or Excellent score. For 74.6 percent of items, all three annotators assign a Good or Excellent score. These results show that most captions in the study reach a strong level of quality.

Figure 3 shows per-lens statistics from the human evaluation study. Caption quality is consistently high, with 88.0% of items rated Good or Excellent by at least two annotators. The emotional lens achieves the highest rating (95.6%) while abstract shows the lowest (85.9%). Inter-annotator agreement is strong across all lenses. In 87.5% of items, at least two annotators agree the caption fits the dataset lens, with literal captions showing highest agreement (94.4%) and abstract the lowest (76.9%). When selecting the best-fitting lens, 96.7% of items have majority agreement and 89.4% match the original dataset category. Among items with unanimous annotator agreement, 99.2% validate the dataset lens assignment. The literal lens demonstrates strongest overall performance while abstract shows lower scores.

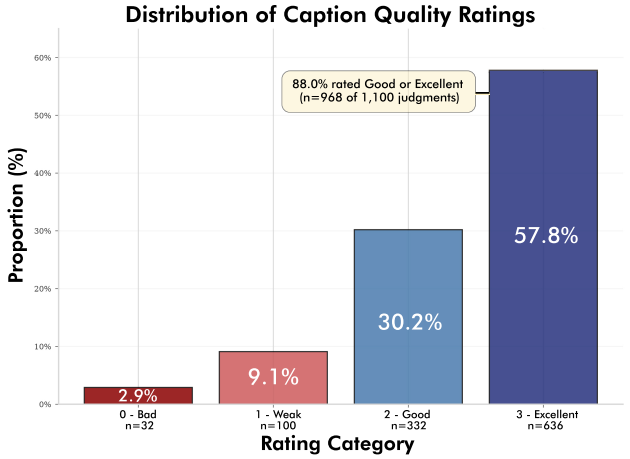

Figure 2. Caption quality ratings over 1,100 valid judgments in the human study.

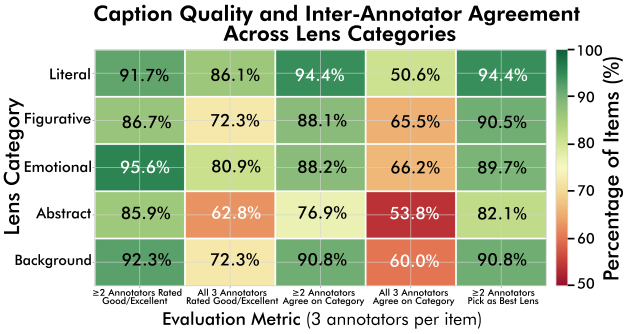

Figure 3. Caption quality and inter-annotator agreement across lens categories. Green indicates higher agreement/quality; Red indicates lower values. Abstract lens shows consistently lower performance.

We measure inter-annotator agreement using Fleiss' κ for items with three annotations (Figure 4). For the four-level caption quality scale, the agreement value is 52.2% with $\kappa = 0.15$, which reflects variation in how annotators judge the difference between Good and Excellent. Grouping the scale into positive and negative ratings increases the agreement value to 83.6% with $\kappa = 0.23$. Lens selection shows stronger consistency. The five-way choice with an extra None option reaches 75.7% agreement with $\kappa = 0.70$, which indicates substantial agreement on the lens that best fits each caption. The Yes or No category question reaches 78.4% agreement with $\kappa = 0.22$, influenced by the large number of Yes responses. Only 4.0% of judgments involve outside information, which suggests that decisions rely on the image, the caption, and the lens definitions. These re-

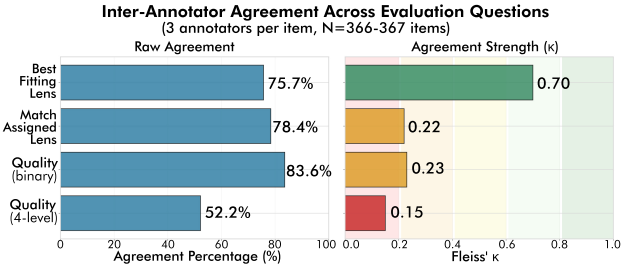

Figure 4. Inter-annotator agreement across three evaluation questions (3 annotators per item). *Agreement* shows percentage of raters choosing the same response within each item. *Fleiss'* κ measures agreement strength: $\kappa < 0.20$ (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), 0.81–1.00 (almost perfect). Best lens selection shows substantial agreement ($\kappa = 0.70$), while quality ratings show high percentage agreement but lower κ due to the 4-level scale and skew toward positive ratings.

sults show that the automatic annotations in Lenses are reliable, with strong caption quality ratings, a clear match between majority lens choice and dataset labels, and substantial agreement on lens selection.

Relation to human preference. While a direct human preference study comparing top-K retrieved results across models would provide the most direct measure of retrieval quality in the non-literal setting, we provide several complementary forms of evidence that approximate human preference. Our human validation study (Sec. 9) confirms that 88.0% of captions are rated Good or Excellent, with substantial inter-annotator agreement on lens assignment ($\kappa = 0.70$), establishing that the underlying annotations align with human judgment. Lens-aware metrics such as LC@10 and All@10 measure whether retrieved results cover multiple valid interpretations of an image, while Lens DCG@10 and Caption DCG@10 assess the ranking quality of those interpretations. Transfer results on ArtEmis further evaluate performance on *human-written* emotional captions, providing a complementary test of alignment with human multi-perspective semantics. Together, these results provide converging evidence that Lenses retrieves results that better reflect the diversity of interpretations people assign to the same image. A controlled human preference study over retrieved results remains a valuable direction for future work.

10. Qualitative examples from Lenses

Figs. 5 and 6 show random examples from the Lenses dataset, where each image is paired with five captions annotated under our Literal, Figurative, Abstract, Emotional, and Background lenses. Fig. 7 shows random examples of image-side prompts, which are used as lens-specific query anchors. Together, these samples illustrate the range of visual domains covered by the dataset, from natural landscapes and abstract art to everyday scenes and human activities, as well as the diversity of non-literal semantics expressed through both captions and image-side prompts. They also provide a qualitative sense of how the same image can support multiple, lens-specific readings that go beyond literal description, further motivating the need for lens-aware retrieval models.

11. Role of Architecture vs. Data Diversity

A natural question is whether the gains of Lenses stem primarily from the diverse multi-lens training data or from the lens-aware architecture itself. To disentangle these factors, we fine-tune CLIP and BGE-VL-MLLM on the same Lenses training split and compare against our full model in Tab. 9.

	I→T		T→I	
	R@1	R@5	R@1	R@5
Zero-shot				
CLIP	8.88	20.97	3.85	16.16
Fine-tuned on Lenses dataset				
CLIP	47.90	77.61	28.07	49.17
BGE-VL-MLLM	80.89	96.16	53.88	77.84
Lenses - All	**89.09**	**98.18**	**58.87**	**80.74**

Table 9. Architecture ablation on the Lenses test set. All fine-tuned models are trained on the same Lenses training split; differences therefore reflect architectural capacity for polysemous alignment.

While Lenses shares the same backbone as BGE-VL-MLLM, it replaces the single pooled embedding with lens-conditioned slots and lens-masked matching, complemented by slot-alignment and diversity losses that prevent embedding collapse. Fine-tuning CLIP on the Lenses data yields only 47.90% / 28.07% R@1 (I→T / T→I), confirming that a single global embedding is insufficient for capturing multi-lens semantics. Lenses outperforms CLIP by +41.19 points on I→T R@1. Even compared to the stronger BGE-VL-MLLM baseline trained on identical data, Lenses improves by +8.2 on I→T and +5.0 on T→I R@1, demonstrating that the lens-aware architecture—not merely data diversity, drives the performance gains. Furthermore, the main paper (Tab. 3) shows that lens-masked matching improves over lens-agnostic matching with embeddings held fixed, isolating the structural contribution of the lens-conditioning mechanism.

12. Prompt design for data generation and validation

We have used a set of text prompts instructions to drive the entire Lenses data pipeline: image selection, image-side prompt generation, category-conditioned captioning, and multi-stage quality control. All prompts share the same five lens definitions, so that supervision is consistent across images, captions, and image-side prompts.

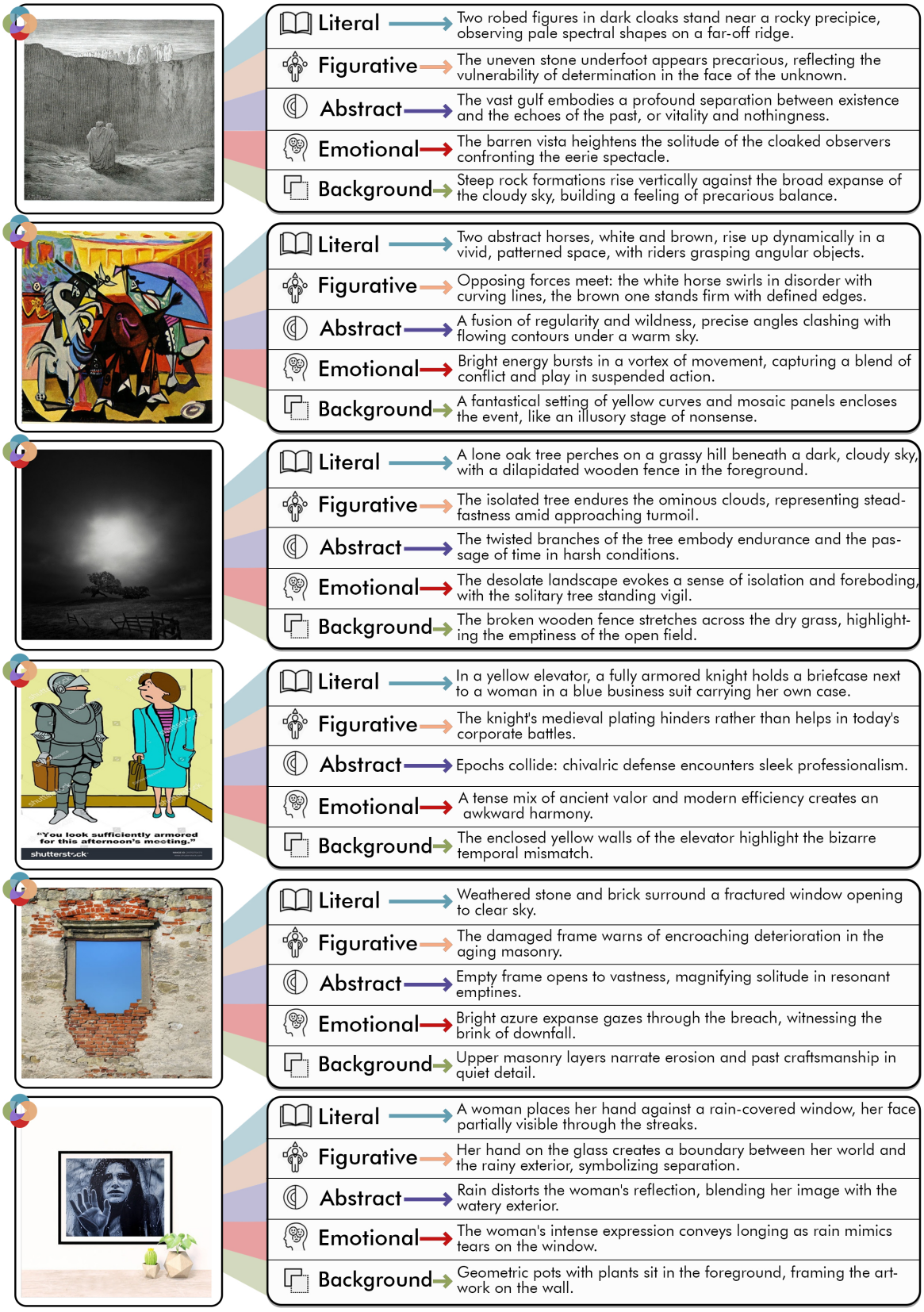

Figure 5. Random caption samples from Lenses.

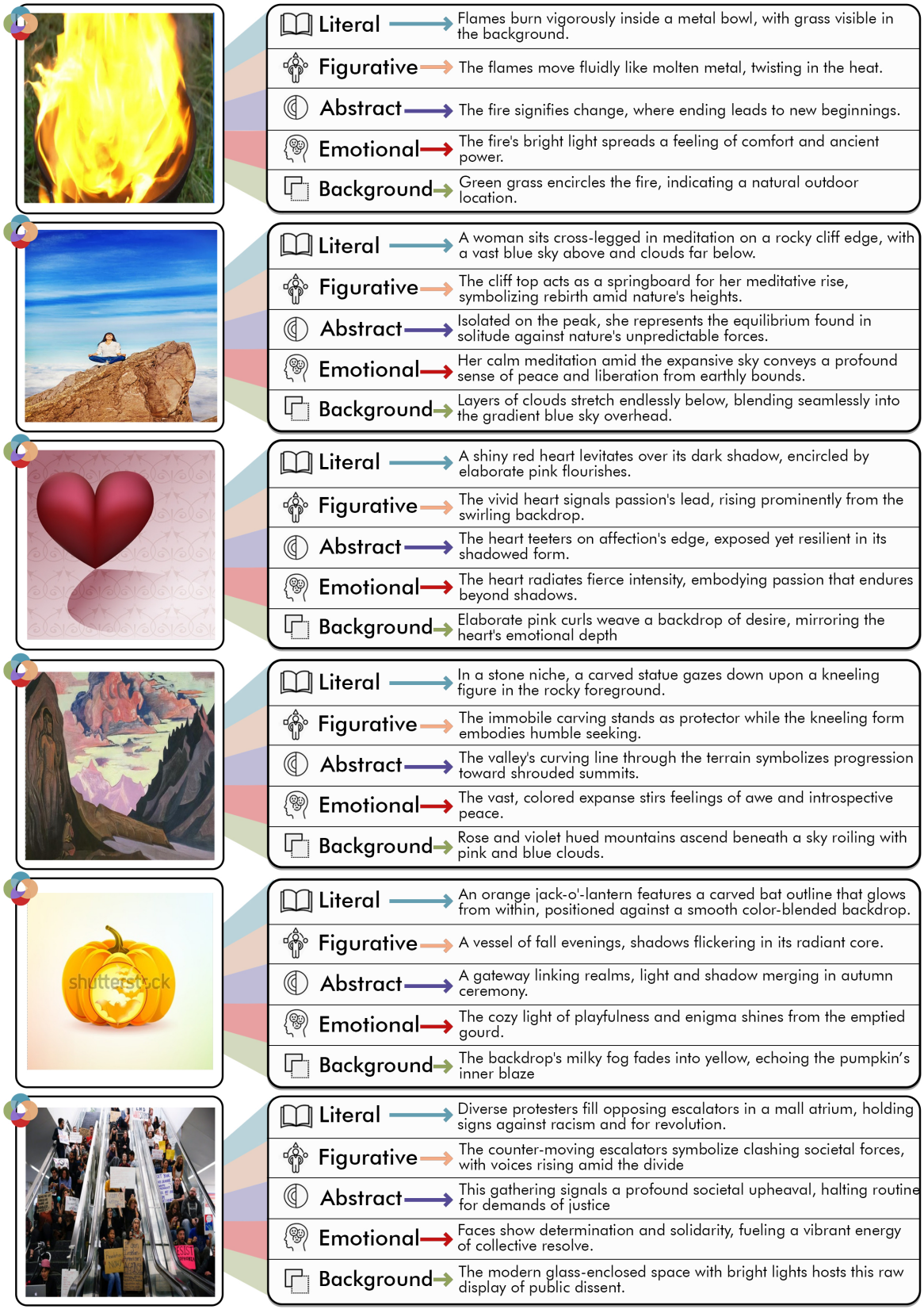

Figure 6. Random caption samples from Lenses.

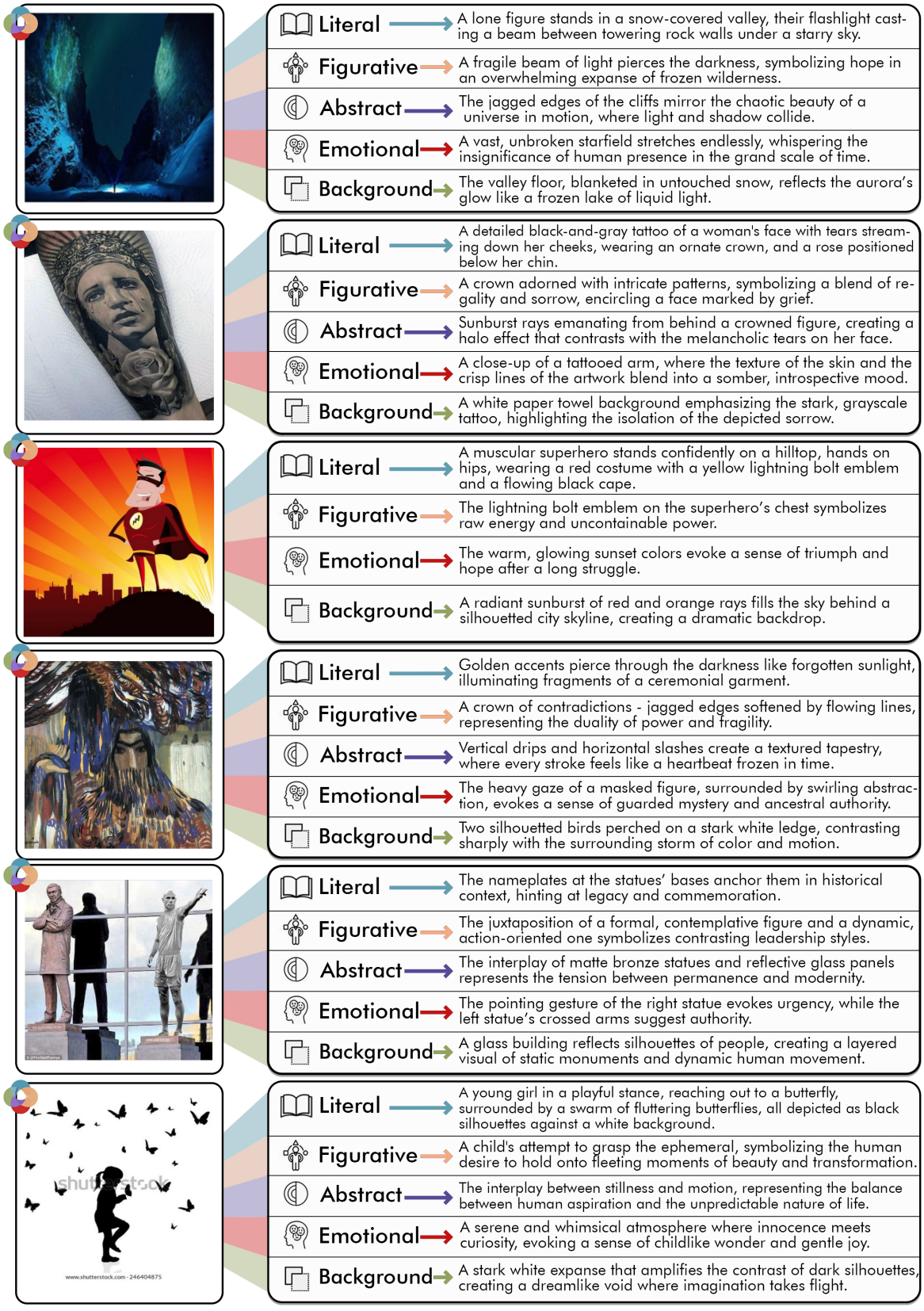

Figure 7. Random image-side prompt samples from Lenses.

Image-level filtering. Before generating annotations, we use Prompt 1 to discard images that do not clearly support multiple non-literal interpretations. Given an image, the model is asked to decide whether it admits *clearly different non-literal interpretations in the Figurative or Abstract categories* and to answer with a single token (`YES`/`NO`). The instructions are intentionally conservative: the model should only return `YES` if it is fully confident that the image supports multiple distinct readings, and it must return `NO` for images that are purely Literal, purely Emotional, or purely Background focused. This stage prunes away uninformative or unambiguous cases so that downstream prompts operate on images with genuine polysemy.

Image-side prompt generation. For images that pass the filter, we generate rich image-side prompts using Prompt 2. The model is instructed to produce a variable-sized set of short prompts, each with a single explicit focus (`"Focus"` field) and an associated `"Category"`. The prompt text must be self-contained, and visually grounded. The template explicitly encourages syntactic variation (imperative, descriptive, contrastive, rhetorical question), and enforces strong lens control:

- Figurative prompts *must* contain at least one idiom drawn from a curated phrase bank of conventional metaphors (*e.g.* *house of cards*, *glass ceiling*, *Trojan horse*), grouped into thematic families (warnings, traps, fragility, complexity, social dynamics, and so on).
- Emotional and Abstract prompts may optionally use an idiom if it fits naturally.
- Literal and Background prompts are prohibited from using phrase bank items.

The phrase bank is a manually curated lexicon of conventional English idioms and fixed multi word expressions (*e.g.* *house of cards*, *elephant in the room*, *black swan*) that encode abstract notions such as evidence, risk, control, and cascading change. We follow standard linguistic treatments of idioms as conventional, often partly non-compositional expressions [4, 9, 10] and design phrase bank as a small, task-specific analogue of idiom resources used in NLP such as the MAGPIE corpus [5].

At the prompt level, we additionally require that roughly 40% of prompts per image use at least one idiom. Internal checks encourage the generator to stop when additional prompts would be redundant, and to avoid hallucinated objects, repeated heads, or near duplicate paraphrases. The resulting JSON records store `Prompt`, `Focus`, and `Category`, which we later reuse as image-side lens prompts during embedding training.

Category-conditioned caption generation. We generate lens-aware captions with Prompt 3. This template asks the model to produce a small set of diverse captions that each present a different way of *seeing* the same image, while maintaining visual grounding. The prompt explicitly requires coverage across all five lenses. As in the image-side prompt template, each caption is accompanied by a `"Focus"` and a `"Category"` field.

Text-only validation and refinement. Both image-side prompts and captions are passed through a text-only cleanup stage using Prompt 4. Here, the model receives a JSON object with a `"Caption"` or `"Prompt"`, its `"Focus"`, and its assigned `"Category"`. The instructions focus on three tasks: lightly edit grammar, spelling, and punctuation and remove machine artifacts (leftover labels, stray quotes, truncation), check that the text is consistent with its Category under the lens definitions, and discard items that cannot be fixed without changing their meaning or hallucinating content. Crucially, the Focus and Category fields are treated as fixed; if they are fundamentally mismatched to the text, the validator is instructed to discard the item by returning an empty string rather than silently reassigning labels. This stage improves fluency and reduces prompt-engineering artifacts before any image-grounded checks.

Test set visual validation. Finally, for all test instances we apply an image-conditioned validator, Prompt 5, which takes the IMAGE together with the JSON array of candidate captions or prompts. This prompt reuses the same lens definitions but now treats the image as the source of truth, explicitly checking both well-formedness and visual grounding of each text. The validator must ensure that every concrete object, attribute, and relation mentioned is supported by the image or can be reliably inferred; otherwise, it may either soften or remove the offending fragment or discard the item entirely. As in the text-only stage, it may not change the Category, and it must discard items that cannot be reconciled with their lens label without altering their core meaning.

Listing 1. PROMPT-Image filtering instruction

```
PROMPT-Image filtering

You are given an image.

Lens definitions:
- Literal: Describes concrete objects and actions
     that are directly visible in the image.
- Figurative: Uses metaphor, idiom, symbolism, or
     other non-literal language where the
    intended
  meaning goes beyond what is literally shown (
    for example, "bad apple", "black box",
  "dirty money").
- Emotional: Focuses on the mood or feeling of
    the scene or characters (for example,
    loneliness,
  joy, tension), not just what things look like.
- Abstract: Talks about high-level ideas or
    themes (for example, freedom, chaos, memory,
```

```
  conflict) instead of concrete objects or
     actions.
- Background: Describes the context, setting,
     environment, style, or artistic technique
     that
  frames the main subject but is not itself the
     main focus, or focuses on background objects.

Task:
1. Decide whether the image supports clearly
     different non-literal interpretations in the
    Figurative or Abstract categories.
2. Answer YES only if you are fully confident
     that the image has multiple distinct possible
    meanings across these non-literal categories.
3. If you are not completely certain that this
     requirement is met, answer NO.
4. Images that are only Literal, only Emotional,
     or only Background focused should also be
    answered with NO.

Output format:
- Do not output captions, categories, or
     explanations.
- Output exactly one word: YES or NO.
```

Listing 2. PROMPT-Image-side prompts instruction

```
PROMPT-Image-side prompts

You are given an IMAGE. Generate a set of text
     PROMPTS that emphasize different
aspects of the scene so that embeddings capture
     varied meanings. Do not fix the
number: produce as many prompts as are genuinely
     distinct, and stop when further
prompts would overlap.

Lens definitions:
- Literal: Describes concrete objects and actions
      that are directly visible in the image.
- Figurative: Uses metaphor, idiom, symbolism, or
      other non-literal language where the
  intended meaning goes beyond what is literally
     shown.
- Emotional: Focuses on the mood or feeling of
     the scene or characters (for example,
  loneliness, joy, tension), not just what things
      look like.
- Abstract: Talks about high-level ideas or
     themes (for example, freedom, chaos, memory,
  conflict) instead of concrete objects or
     actions.
- Background: Describes the context, setting,
     environment, style, or artistic technique
  that frames the main subject but is not itself
     the main focus, or focuses on background
  objects.

RULES
- One focus per prompt (object, region, relation,
      texture, color palette, composition cue,
  or theme).
- Each prompt is self-contained (no "this image"
     or "this photo"), with 10-28 words.
- Rotate syntax (imperative, descriptive,
     contrastive clause, rhetorical question) to
  increase variety.
- Use the Categories when applicable: Literal,
     Figurative, Emotional, Abstract, Background.

CATEGORIES
- Literal: straightforward, denotative
     descriptions grounded in visible content.
- Figurative: MUST use at least one phrase from
     PHRASE_BANK.
- Emotional: focuses on atmosphere or affect.
- Abstract: treats the scene as an idea or theme.
- Background: reads the context independent of
     the main subject.

FIGURATIVE USE
- Figurative prompts MUST include at least one
     idiom from PHRASE_BANK (conventional
  meaning only).
- Emotional or Abstract prompts MAY include an
     idiom if it fits naturally.
- Literal or Background prompts MUST NOT contain
     idioms from PHRASE_BANK.
- For this image, target about 40 percent of
     prompts to be idiomatic (round up); avoid
  reusing the same idiom twice when possible.

PHRASE_BANK

Warnings & Evidence
smoking gun; canary in a coal mine; third rail;
     red flag; bellwether; tipping point;
line in the sand; watershed moment; writing on
     the wall

Traps, Dilemmas & Costly Wins
Catch-22; Hobson's choice; Faustian bargain;
     poisoned chalice; Pyrrhic victory;
double-edged sword; pick your poison; rock and a
     hard place

Deception, Illusion & Soft Power
smoke and mirrors; shell game; red herring;
     Trojan horse; paper tiger; Potemkin village;
bait and switch; dog whistle; stalking horse

Fragility, Risk & Volatility
house of cards; sword of Damocles; thin ice;
     tightrope; powder keg; shaky ground;
glass jaw; walking a minefield

Effort, Struggle & Limits
Sisyphean task; Herculean task; uphill battle;
     swimming upstream; long slog;
running on fumes; burning the candle at both ends
     ; move the needle

Complexity, Change & Cascades
Gordian knot; rabbit hole; butterfly effect;
     domino effect; sea change; paradigm shift;
tectonic shift; ripple effect

Time, Deadlines & Inevitability
eleventh hour; last straw; point of no return;
     train has left the station;
on borrowed time; wheels already in motion; the
     clock is ticking
```

```
Boundaries, Access & Control
velvet rope; walled garden; back door; open
    secret; revolving door; gatekeeper;
skeleton key; closed book

Rarity, Ambition & Extremes
black swan; once in a blue moon; lightning in a
    bottle; needle in a haystack;
moonshot; long tail; outlier; edge case

Social Dynamics & Discourse
elephant in the room; echo chamber; straw man;
    moving the goalposts; glass ceiling;
party line; chorus of approval; whisper network

Quests, Guides & Obsessions
holy grail; white whale; north star; lodestar;
    Rosetta stone; golden fleece;
compass point; guiding light

Consequences & Endings
nail in the coffin; slippery slope; pay the piper
    ; burn bridges; day of reckoning;
tipping the scales; crying over spilled milk;
    pandora's box (opened)

Systems, Tech & Operations (modern idioms and
    jargon widely used metaphorically)
single point of failure; bus factor; happy path;
    garbage in, garbage out;
canary release; fail safe; black box; feedback
    loop

Perspective, Mirrors & Multiplicity
house of mirrors; two sides of the same coin;
    looking glass world; cat's eye view;
blind spot; bird's eye view; tunnel vision;
    double take

Renewal, Cycles & Resilience
phoenix rising; second wind; turning the page;
    fresh start; green shoots;
silver lining; tide turns; springboard

Add ons
tip of the iceberg; calm before the storm;
    stacked deck; zero-sum game; moving target;
shifting sands; under the radar; in plain sight;
    back to the drawing board; trial by fire;
silver bullet; technical debt; red tape; cut
    corners; level playing field; break the ice;
open the floodgates; paper trail; wild card; ace
    up the sleeve

OUTPUT (JSON array only)
[
  {"Prompt":"<text>","Focus":"<object/region/idea
    >",
   "Category":"Literal|Figurative|Emotional|
    Abstract|Background"},
  ...
]

INTERNAL CHECK (do not print)
- Variable set size with clear novelty; stop
    before redundancy.
- About 40 percent of prompts for this image
    contain PHRASE_BANK idioms; none in Literal
  or Background.
- Strong visual grounding; no hallucinated
    objects; minimal head noun or idiom
    repetition.
```

Listing 3. PROMPT-Category-conditioned caption generation instruction

```
PROMPT-Category-conditioned caption generation

You are given an IMAGE. Generate a set of diverse
    , self-contained captions that each
present a distinct way of "seeing" the same
    picture. Maintain visual grounding while
weaving in meaning-dense idioms, metaphors, and
    loaded phrases when appropriate.

Lens definitions:
- Literal: Describes concrete objects and actions
     that are directly visible in the image.
- Figurative: Uses metaphor, idiom, symbolism, or
     non-literal language where the intended
  meaning goes beyond what is literally shown.
- Emotional: Focuses on the mood or feeling of
    the scene or characters (for example,
  loneliness, joy, tension), not just what things
     look like.
- Abstract: Talks about high-level ideas or
    themes (for example, freedom, chaos, memory,
  conflict) instead of concrete objects or
    actions.
- Background: Describes the context, setting,
    environment, style, or artistic technique
  that frames the main subject but is not itself
    the main focus, or focuses on background
  objects.

GOALS
- Coverage: span literal description, mood,
    symbolism, setting, and style.
- Figurative control: add idioms or metaphors
    only when they naturally fit visible elements
    .
- Non-overlap: each caption must highlight a
    different facet (no near duplicates).

MANDATES ABOUT PHRASE_BANK
- For every Figurative caption, you MUST include
    at least one phrase from PHRASE_BANK and
  use it in its conventional sense.
- Emotional or Abstract captions MAY include a
    PHRASE_BANK item if it naturally fits.
- Literal and Background captions SHOULD NOT use
    PHRASE_BANK items.
- Across the whole set for this image, ensure at
    least 40 percent of captions use at least
  one PHRASE_BANK item.

METHOD (think, then write)
1) Parse the scene: salient objects, relations,
    composition cues (rule of thirds, symmetry
   or leading lines), materials or textures,
    colors, foreground or background, and any
   actions.
2) Map visible concretes to abstract themes (risk
    , fragility, power, isolation, renewal,
```

```
   deception, scarcity, etc.). Consider candidate
     idioms whose meanings match these themes.
3) For each caption, pick ONE focus and ONE
    rhetorical device set (if any). Keep it
    concise
   (8-22 words), standalone (no "this image" or "
    this photo"), and free of brand names or
   private data.

CATEGORIES (produce at least one caption per
    category)
- Literal: straightforward, denotative
    description grounded in visible content.
- Figurative: MUST use at least one phrase from
    PHRASE_BANK.
- Emotional: emphasizes atmosphere or affect.
- Abstract: presents the scene as an idea or
    theme.
- Background: reads the context independently of
    the main subject.

STYLE GUARDRAILS
- Faithfulness: no hallucinated objects;
    figurative language must align with what is
    visible.
- Diversity: vary syntax and vocabulary; avoid
    repeating the same idiom family or head nouns
    .
- Safety: avoid stereotypes and outdated or
    offensive expressions.

PHRASE_BANK (curated; choose zero, one, or two
    per caption as allowed above)

Warnings & Evidence
smoking gun; canary in a coal mine; third rail;
    red flag; bellwether; tipping point;
line in the sand; watershed moment; writing on
    the wall

Traps, Dilemmas & Costly Wins
Catch-22; Hobson's choice; Faustian bargain;
    poisoned chalice; Pyrrhic victory;
double-edged sword; pick your poison; rock and a
    hard place

Deception, Illusion & Soft Power
smoke and mirrors; shell game; red herring;
    Trojan horse; paper tiger; Potemkin village;
bait and switch; dog whistle; stalking horse

Fragility, Risk & Volatility
house of cards; sword of Damocles; thin ice;
    tightrope; powder keg; shaky ground;
glass jaw; walking a minefield

Effort, Struggle & Limits
Sisyphean task; Herculean task; uphill battle;
    swimming upstream; long slog;
running on fumes; burning the candle at both ends
    ; move the needle

Complexity, Change & Cascades
Gordian knot; rabbit hole; butterfly effect;
    domino effect; sea change; paradigm shift;
tectonic shift; ripple effect

Time, Deadlines & Inevitability
eleventh hour; last straw; point of no return;
    train has left the station;
on borrowed time; wheels already in motion; the
    clock is ticking

Boundaries, Access & Control
velvet rope; walled garden; back door; open
    secret; revolving door; gatekeeper;
skeleton key; closed book

Rarity, Ambition & Extremes
black swan; once in a blue moon; lightning in a
    bottle; needle in a haystack;
moonshot; long tail; outlier; edge case

Social Dynamics & Discourse
elephant in the room; echo chamber; straw man;
    moving the goalposts; glass ceiling;
party line; chorus of approval; whisper network

Quests, Guides & Obsessions
holy grail; white whale; north star; lodestar;
    Rosetta stone; golden fleece;
compass point; guiding light

Consequences & Endings
nail in the coffin; slippery slope; pay the piper
    ; burn bridges; day of reckoning;
tipping the scales; crying over spilled milk;
    pandora's box (opened)

Systems, Tech & Operations (modern idioms and
    jargon widely used metaphorically)
single point of failure; bus factor; happy path;
    garbage in, garbage out; canary release;
fail safe; black box; feedback loop

Perspective, Mirrors & Multiplicity
house of mirrors; two sides of the same coin;
    looking glass world; cat's eye view;
blind spot; bird's eye view; tunnel vision;
    double take

Renewal, Cycles & Resilience
phoenix rising; second wind; turning the page;
    fresh start; green shoots; silver lining;
tide turns; springboard

OUTPUT (JSON array only; no extra text)
[
  {
    "Caption": "<text>",
    "Focus": "<object/region/idea>",
    "Category": "Literal|Figurative|Emotional|
    Abstract|Background"
  },
  ...
]

QUALITY CHECK (internal; do not print your notes)
- If an idiom would mislead (for example, "
    smoking gun" without any evidence motif),
  choose a different phrase.
- Ensure at least 70 percent token difference
    between captions; vary heads and modifiers.
- Prefer concrete to abstract metaphors; keep
```

```
    meanings conventional and consistent with the
  visual evidence.
```

Listing 4. PROMPT-Validation and refinement instruction

```
PROMPT-Validation and refinement

You are a validator and light editor for captions
     and prompts in a vision-language dataset.

Each input item is a JSON object:
{
  "Caption" | "Prompt": "<candidate text>",
  "Focus": "<object/region/idea>",
  "Category": "Literal|Figurative|Emotional|
    Abstract|Background"
}

Your job is to validate and, when needed,
    minimally revise each item so that texts are
well formed, free of artifacts, and consistent
    with their assigned Category.

Lens definitions (for Category):
- Literal: Describes concrete objects and actions
     that are directly visible in the image.
- Figurative: Uses metaphor, idiom, symbolism, or
     non-literal language where the intended
  meaning goes beyond what is literally shown.
- Emotional: Focuses on the mood or feeling of
    the scene or characters (for example,
  loneliness, joy, tension), not just what things
     look like.
- Abstract: Talks about high-level ideas or
    themes (for example, freedom, chaos, memory,
  conflict) instead of concrete objects or
    actions.
- Background: Describes the context, setting,
    environment, style, or artistic technique
  that frames the main subject but is not itself
    the main focus, or focuses on background
  objects.

Validation rules:
1. Grammar and fluency
   - Fix grammar, spelling, and punctuation.
   - Remove machine artifacts (leftover labels,
    extra quotes, truncation, placeholders).
   - Keep the original meaning and perspective as
     much as possible.

2. Category alignment
   - Check that the text matches its Category
    according to the definitions above.
   - Do not change Focus or Category. If they
    clearly cannot match, discard the item.

3. Discard rule
   - If a text cannot be fixed without changing
    its meaning or hallucinating content,
     discard it by outputting an empty string for
     the Caption/Prompt value.

OUTPUT (JSON array only; no extra text)
[
  {
    "Caption" | "Prompt": "<final text or \"\" if
     discarded>",
    "Focus": "<object/region/idea>",
    "Category": "Literal|Figurative|Emotional|
    Abstract|Background"
  },
  ...
]
```

Listing 5. PROMPT-Test set validation instruction

```
PROMPT-Test set validation

You are a visual grounding validator and light
    editor for test set captions and prompts
in a vision-language dataset, using the image as
    the source of truth.

For each test instance you receive:
- One IMAGE.
- A JSON array of items of the form:
  {
    "Caption" | "Prompt": "<candidate text>",
    "Focus": "<object/region/idea>",
    "Category": "Literal|Figurative|Emotional|
    Abstract|Background"
  }

Your job is to check that every text is:
1) well formed and free of obvious artifacts, and
2) visually grounded in the IMAGE and consistent
    with its Category.

If an item fails these checks and cannot be fixed
     with small edits, you must discard it.

Lens definitions (for Category):
- Literal: Describes concrete objects and actions
     that are directly visible in the image.
- Figurative: Uses metaphor, idiom, symbolism, or
     other non-literal language where the
  intended meaning goes beyond what is literally
    shown, while still being conceptually
  connected to the visible content.
- Emotional: Focuses on the mood or feeling of
    the scene or characters (for example,
  loneliness, joy, tension), not just what things
     look like.
- Abstract: Talks about high-level ideas or
    themes (for example, freedom, chaos, memory,
  conflict) instead of concrete objects or
    actions, but still rooted in what the image
  suggests.
- Background: Describes the context, setting,
    environment, style, or artistic technique
  that frames the main subject but is not itself
    the main focus, or focuses on background
  objects.

Verification rules:

1. Grammar and artifacts
   - Fix grammar, spelling, and punctuation when
    needed.
   - Remove machine artifacts such as leftover
    labels ("Caption:", "Category:", indices),
     stray quotes, placeholders ("<text>", "<
    object>"), or truncated fragments.
```

```
   - Keep the original intention and perspective
     as much as possible.

2. Visual grounding
   - Check that all concrete objects, attributes,
     actions, and relations mentioned in the
     text are supported by the IMAGE.
   - If a specific object, detail, or relation is
     not visible or cannot be reliably
     inferred from the IMAGE, either:
       * remove or soften that part of the text
     while keeping the rest grounded, or
       * if that would destroy the main meaning,
     discard the item.
   - Do not introduce new objects, actions, or
     specific details that are not seen
     in the IMAGE.

3. Category alignment
   - Check that the text matches its Category
     according to the lens definitions.
   - Do not change Category or Focus. Instead:
       * If the text can be edited slightly so
     that it matches its Category and remains
          visually grounded, perform the edit.
       * If the text clearly cannot match the
     Category without changing its core meaning,
          discard the item (set the text to an
     empty string).

4. Figurative and idiomatic content
   - Figurative items may use idioms or metaphors
     , but they must remain conceptually
     compatible with the IMAGE (no idioms that
     imply events or evidence that are clearly
     absent).
   - Do not add new idioms. You may remove or
     simplify them if they conflict with the
     visual content.

Discard rule:
- If a text is severely ungrammatical, dominated
    by artifacts, clearly ungrounded in the
  IMAGE, or fundamentally mismatched with its
    Category in a way that cannot be fixed by
  small edits, discard it by setting its Caption/
    Prompt value to "".

OUTPUT (JSON array only; no extra text)
Return a JSON array with the same structure as
    the input, where each item is:
{
  "Caption" | "Prompt": "<final text or \"\" if
    discarded>",
  "Focus": "<object/region/idea>",
  "Category": "Literal|Figurative|Emotional|
    Abstract|Background"
}

- If you apply minor corrections, output the
    corrected text.
- If you decide to discard an item, output an
    empty string "" for the Caption/Prompt value.
```

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] 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. 3

[3] 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

[4] Sam Glucksberg and Matthew S McGlone. *Understanding figurative language: From metaphor to idioms*. Number 36. Oxford University Press, 2001. 11

[5] Hessel Haagsma, Johan Bos, and Malvina Nissim. Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, 2020. 11

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1 (2):3, 2022. 4

[7] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2

[8] 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

[9] Rosamund Moon. *Fixed Expressions and Idion1s in English: A Corpus-Based Approach*. Oxford University Press, 1998. 11

[10] Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. Idioms. *Language*, 70(3):491–538, 1994. 11

[11] 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

[12] 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. 1

[13] 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

[14] xAI. Grok 4.1, 2025. Large language model developed by xAI. 3

[15] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[16] 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

[17] 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. 2

[18] 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. 1, 2, 4