AutoV: Loss-Oriented Ranking for Visual Prompt Retrieval in LVLMs

Yuan Zhang[1,2], Chun-Kai Fan[1], Sicheng Yu[2], Junwen Pan[2], Tao Huang[3], Ming Lu[1], Kuan Cheng[1], Qi She[2], and Shanghang Zhang[1†]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2] ByteDance
[3] Shanghai Jiao Tong University

Code https://github.com/Gumpest/AutoV

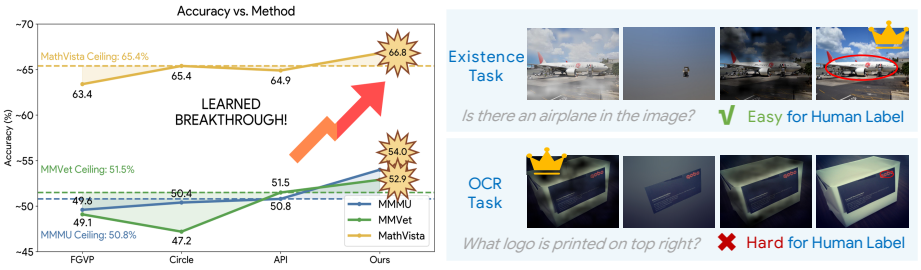
Abstract. Inspired by text prompts in large language models, visual prompts have been explored to enhance the perceptual capabilities of large vision-language models (LVLMs). However, performance tends to saturate under single visual prompt designs, making further prompt engineering increasingly ineffective. To address this limitation, we shift from prompt engineering to prompt retrieval and propose AutoV, a lightweight framework for instance-adaptive visual prompt identification. Given an input image and a textual query, AutoV automatically locates the most suitable visual prompt from a diverse candidate pool. Training such a retrieval framework requires prompt-level supervision, yet prompt quality is inherently ambiguous and difficult to assess reliably, even for humans. To enable automatic supervision, we evaluate visual prompts using a pre-trained LVLM and label them according to their prediction losses. Using the loss-oriented ranking as a robust training signal, AutoV learns to retrieve the query-aware optimal prompt for each instance without manual annotation. Experiments indicate that AutoV enhances the performance of various LVLMs on image understanding, captioning, grounding, and classification tasks. For example, AutoV improves LLaVA-OV by **10.2**% on VizWiz and boosts Qwen2.5-VL by **3.8**% on MMMU, respectively.

Keywords: Visual Prompt · Retrieval · Loss-Oriented · LVLMs

1 Introduction

Recent advancements in large language models (LLMs) [1,3,5,8,10,35,47] have spurred significant progress in large vision-language models (LVLMs), which connect visual encoders with language model decoders to effectively adapt textual reasoning capabilities to visual understanding tasks [4,15,23,26,31,43,44]. Consequently, visual input plays a crucial role in LVLMs, enabling precise perception and fine-grained grounding across various multimodal scenarios.

† Corresponding author.



(a) Performance Saturation. **(b) Visual Prompt Labeling Difficulty.**

Fig. 1: Motivation of AutoV. (a) Performance saturation. Existing visual prompts approach benchmark ceilings, limiting further gains from prompt engineering. (b) Labeling difficulty and task diversity. Optimal prompts vary across tasks, and the crown denotes the one leading to the correct answer. While the optimal prompt is easy to identify in the top example, it is much harder to determine in the bottom one.

Inspired by textual prompting in LLMs [25, 49, 67], visual prompting has emerged as an effective paradigm for guiding LVLM attention [30, 42, 51, 62]. By injecting structured visual cues, such as blur masks, bounding circles, or attention heatmaps, visual prompts encourage models to focus on task-relevant regions. However, existing approaches predominantly rely on fixed, heuristically designed prompts. While these handcrafted strategies can yield improvements on certain benchmarks, they implicitly assume that a single prompt design generalizes across diverse visual scenes and textual queries.

In practice, this assumption rarely holds. As shown in Figure 1(a), heuristic prompts tend to approach benchmark-specific performance ceilings, leaving limited room for improvement through prompt engineering alone. Moreover, prompt effectiveness varies across tasks and instances. For example, in Figure 1(b), attention-based masks may benefit OCR-sensitive tasks, while highlight circles emphasize salient objects for detection. These observations suggest that **the optimal visual prompt is inherently instance-dependent**, making fixed prompt designs fundamentally limited. This insight motivates a paradigm shift from prompt engineering to prompt retrieval: instead of searching for a universally optimal visual prompt, we ask a more flexible question—given a specific image-query pair, which visual prompt is most suitable?

To this end, we introduce **AutoV**. This innovative visual-prompting retrieval framework dynamically selects the optimal visual prompt from a set of candidates, explicitly tailored to the textual query and the input image. Specifically, we first generate various visual prompt candidates and encode their representations. This diversity enables the framework to capture distinct visual representations that may be optimal for various query-image pairs. Next, we design a lightweight ranking network that efficiently integrates visual and textual information to predict the ranking preferences over prompt candidates. Finally, the highest-ranking candidate is chosen as the input for LVLMs during inference.

A central challenge in learning such a retrieval framework lies in supervision. As illustrated in Figure 1(b), the quality of visual prompts is often

ambiguous and task-dependent, making reliable human annotation difficult, particularly for OCR or reasoning-heavy tasks. To address this issue, we introduce a fully automated supervision strategy: each prompt candidate is evaluated by a pre-trained LVLM and ranked according to its prediction loss. This loss-oriented ranking provides a stable training signal, enabling the ranking network to learn query-aware prompt preferences without manual annotation.

Extensive experiments demonstrate that AutoV consistently enhances a wide range of LVLMs across diverse tasks, including image understanding, captioning, grounding, and classification. For example, AutoV improves LLaVA-OneVision by **10.2**% on VizWiz and boosts Qwen2.5-VL by **3.8**% on MMMU. Beyond benchmark gains, AutoV generalizes robustly across closed-source models and integrates seamlessly without additional fine-tuning of the backbone LVLM.

In summary, our contributions are threefold:

- We introduce AutoV, a novel automatic visual prompting framework for large vision-language models that adaptively retrieves the optimal visual prompt from a diverse candidate pool, explicitly tailored to textual query.
- We create a scalable and totally automated data collection pipeline equipped with a reward-based loss to effectively train a lightweight ranking network, facilitating accurate and adaptive visual prompt selection without the need for extensive manual annotations.
- We perform comprehensive experiments to validate the efficacy and robustness of AutoV across tasks, with consistent and significant improvements over existing visual prompt engineering. Once trained, AutoV integrates seamlessly into various LVLMs without additional fine-tuning.

2 Related Work

2.1 Large Vision-Language Models

Recent advances in large language models [1, 8, 10, 35, 39, 47] have catalyzed a new wave of large vision-language models [2, 4, 15, 27, 28, 31]. By coupling a high-capacity visual encoder with an LLM decoder, LVLMs transfer textual reasoning to visual understanding, and the quality of input images determines the lower bound of multimodal abilities. For instance, the popular LLaVA-OneVision [26] and Qwen2.5-VL [5] adopt a higher and dynamic resolution recipe for training to enhance their fine-grained reasoning, while vision-centric scaling in InternVL2 [11] enlarges the ViT [17] backbone to boost grounding capability. However, the above methods require training times of several hundred GPU days. Here, **can we further boost the existing LVLMs by optimizing the visual input itself during the inference stage?** The answer is **Yes**. Visual prompting [24, 30, 40, 42, 55, 57, 65] is an emerging and effective technique in this direction.

2.2 Visual Prompting for LVLMs

LVLMs relying solely on textual inputs exhibited two key limitations: visual hallucinations [6, 22] and language-driven biases [37, 50]. To better mitigate these

challenges, visual prompting [30, 42, 51, 62] has emerged as an alternative way of supplying cues in the visual modality. Compared with textual prompts, it is more concise and explicit, while allowing guidance that reaches fine-grained, pixel-level details. Early work [40] shows that simply drawing a red circle around a target object steers CLIP [38] attention, yielding great performance on referring-expression and keypoint-localization benchmarks. FGVP [55] extends this idea with fine-grained prompting, replacing coarse shapes with blur-reversal prompts based on segmentation masks, further improving accuracy on RefCOCO [63]. AlphaCLIP [42] trains an augmented version of CLIP with an auxiliary alpha channel for indicating regions at the expense of increased training overhead. API [57] proposes attention prompting on images, overlaying text-guided attention heatmaps. Although previous studies have explored various visual prompts, the marginal gains brought by such engineering designs are gradually diminishing. To address this limitation, we propose a visual prompt retrieval mechanism that integrates existing prompt designs, exploiting their complementary advantages.

2.3 Visual Retrieval for LVLMs

Existing visual prompt retrieval pipelines generally adhere to three main paradigms. (1) The heuristic paradigm identifies benchmark-level optimal parameters through extensive evaluation without instance-specific adaptation. (2) The random paradigm offers a baseline characterized by instance-level variation. (3) The Mixture-of-Experts (MoE) paradigm [7, 18, 68] provides the most direct form of adaptation by utilizing GateNet specifically for visual retrieval. Notably, AutoV differs fundamentally from retrieval-augmented methods [36, 54]. Instead of retrieving external visual data with additional finetuning, AutoV performs lightweight in-model prompt selection based solely on the input image features within a compact candidate pool, resulting in significantly improved efficiency.

3 Proposed Approach: AutoV

In this section, we propose the AutoV for optimal visual prompt retrieval. Our framework includes four critical components: (1) feature extraction of prompt candidates, (2) candidate ranking consisting of modality interaction and single-modality mapping modules, (3) reward loss supervision along with automated data generation pipeline, and (4) an efficient and robust inference pipeline. The overall architecture is illustrated in Figure 2, where the four modules are color-coded for clarity. Best viewed in color.

3.1 Feature Extraction of Prompt Candidates

The primary goal is to effectively choose the most appropriate visual prompt in response to various text queries and input images. For a group of n input visual

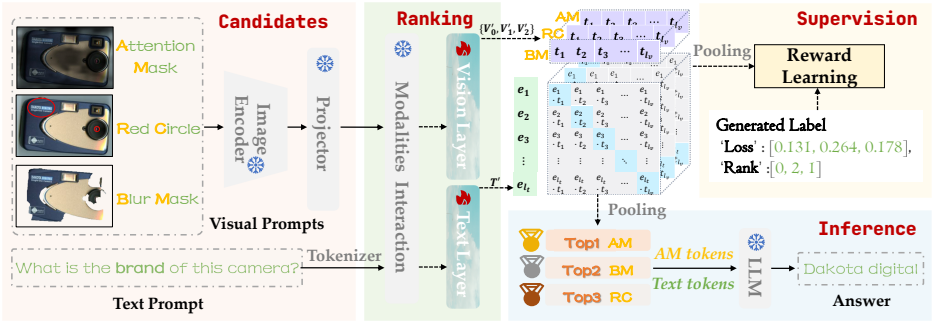


Fig. 2: The illustration of AutoV. It comprises four key components: representation extraction from candidates, ranking network for reward score, reward-supervised training, and inference. It retrieves the prompt tailored to each query-image pair.

prompts $\{x_1, x_2, ..., x_n\}$, we employ a visual encoder $g(.)$ (*e.g.*, like CLIP [38]) to generate the visual feature:

$$\boldsymbol{Z}_i := g(x_i) \text{ with } i \in \{1, ..., n\}. \tag{1}$$

Next, we apply a projection matrix $\boldsymbol{W}$ to transform each visual feature $\boldsymbol{Z}_i$ into language embedding tokens $\boldsymbol{V}_i = \boldsymbol{W} \times \boldsymbol{Z}_i$, aligning the same dimensionality as the word embedding space in the language model. As a result, we obtain a set of visual candidates represented in the form of visual tokens $\boldsymbol{V}_i \in \mathbb{R}^{L \times D}$ to rank, where L denotes the length of vision tokens and D is the language embedding dimension. Notably, both the visual encoder and the alignment projector are directly inherited from a fully trained LVLM, and we adopt them without any additional fine-tuning.

3.2 Candidate Ranking Network

The goal of the ranking network is to effectively rank the visual prompt candidates based on their relevance to the textual query, by modeling the relationship between the visual tokens and the input text. Therefore, we first propose the modality interaction module to enable visual tokens to fuse the context information of text tokens. Inspired by observations in [64] that modality-fusion occurs in the early layers of the LLM, we use the first layer of the LLM decoder $f(.)$ for modality interaction. Thus, visual candidate tokens $\boldsymbol{V}_i$ and text tokens $\boldsymbol{T}$ are concatenated and fed into the interaction module:

$$\boldsymbol{H}_i := f(\text{concat}(\boldsymbol{V}_i, \boldsymbol{T})), \tag{2}$$

with the outputs are $\widetilde{\boldsymbol{V}}_i = \boldsymbol{H}_i[: l_v]$ and $\widetilde{\mathbf{T}} = \boldsymbol{H}_i[-l_t :]$. Next, we input visual candidate tokens $\widetilde{\boldsymbol{V}}_i$ into a vision mapping module, which is implemented using a feed-forward network (FFN). The generated final candidate features are $\boldsymbol{V}_i^{'} \in \mathbb{R}^{l_v \times h}$, where $h \ll D$ is the output dimension of the vision mapping module to

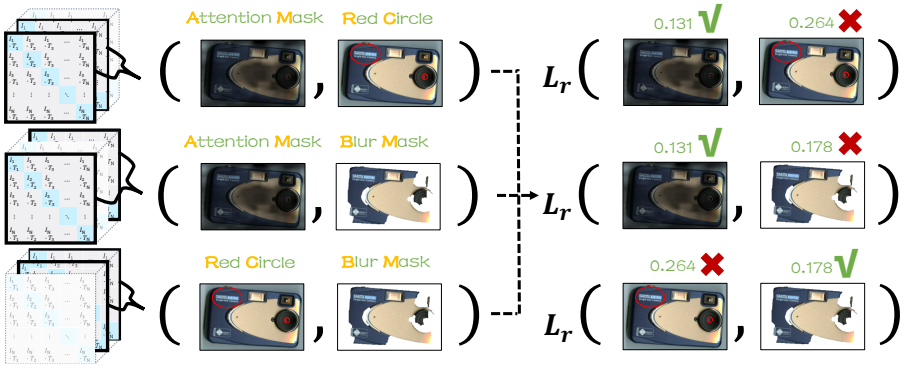


Fig. 3: The reward loss of AutoV. As a concrete example, we illustrate the **pairwise** combination process using the visual prompts from Figure 2, demonstrating how our supervision signal is applied.

save computational cost. Besides, the query tokens $\widetilde{\mathbf{T}}$ are also converted by a text mapping module to get the query embeddings $\boldsymbol{T}^{'} \in \mathbb{R}^{l_t \times h}$. In AutoV, we define visual retrieval as a customized ranking task rather than a classification or regression problem, and the reason is discussed in Section 3.3. Therefore, we calculate the reward scalar for given visual candidate tokens $\boldsymbol{V}_i^{'}$, where the context similarity with the query serves as the quantitative target. We utilize a cross-attention operation to achieve:

$$s(\mathrm{VP}_i) := \mathrm{mean}(\text{cross-attn}(\boldsymbol{V}_i^{'}, \boldsymbol{T}^{'})), \tag{3}$$

where VP_i denotes the ith visual prompt candidate. Here, only the vision and text mapping modules need additional training, each consisting of a single FFN with two linear layers. The modules comprise $2h(D+h+2)$ parameters in total, thereby resulting in lower training costs.

3.3 Reward Loss Supervision

Automated Data Generation. Training a retrieval model requires prompt-level supervision, yet the quality of a visual prompt is inherently ambiguous and lacks a reliable evaluation criterion, **even for humans**. To address this, we construct supervision signals directly from the intrinsic behavior of a pre-trained LVLM. Our approach is grounded in a simple intuition: *a superior visual prompt should induce lower conditional language modeling loss given the visual-question pair.* Rather than relying on one-hot correctness signals, which may reflect the LVLM's language priors instead of true visual grounding [53,58], we optimize the ranking network using relative losses. In this way, the loss serves as a calibrated proxy for model confidence, where *lower loss implies stronger alignment between the prompt and the input instance* [20, 66].

For each group of n visual candidates, we generate n (*VP*, *query*) pairs and comprehensively evaluate them. Each pair is processed by a fully trained vision-language model to compute its *combination loss*, defined as the modeling loss

relative to the ground truth, and sorted to get a rank. To ensure high-quality ranking data, we implement two filtering criteria for (*VP*, *query*) pairs: (1) we remove pairs exhibiting low loss variance, as their insensitivity to visual prompts suggests answers are generated from the language priors rather than actual input content; (2) we exclude pairs with excessively high average loss, which indicates outlier case or weak visual prompt-query correlation.

For each raw (*image*, *query*) pair, we generate a training sample structured as a ***quadruple***:

$$< query, \{vp_0, vp_1, ..., vp_{n-1}\}, rank, reward\ loss >,$$

with an illustrative example provided in Figure 2, and the dataset we adopt is detailed in Appendix A.

Reward Loss Supervision. Inspired by the reward modeling from human feedback in OpenAI [34], we train our ranking network using reward modeling to align its partial order with the optimal visual prompt. Here, we model visual retrieval as a customized ranking task rather than a classification or regression problem. This is because the objective is not to assign absolute labels or scores, but to identify the most appropriate prompt from a set of candidates by comparing their relative effectiveness in enhancing multimodal understanding. Ranking allows the model to learn fine-grained preferences across diverse query-image combinations and better align with the real-world setting.

In Figure 3, we pair all n visual prompt candidates exhaustively, ensuring each prompt is combined with every other prompt exactly once. This results in generating $\mathcal{C}(n, 2)$ training pairs for each instance. The reason is that, unlike pointwise scoring, which assigns low scores to all responses for inherently challenging tasks, pairwise comparison allows for finer discrimination between responses. Based on the rank and loss from the quadruples, we reinforce the visual prompt with lower loss (chosen sample) for each pair, while penalizing the higher-loss visual prompt (rejected sample).

Specifically, the reward loss $\mathcal{L}_r$ is designed as:

$$\mathcal{L}_r := -\frac{1}{\mathcal{C}(n,2)}\mathbb{E}\left[\log\left(\sigma\big(s(\mathrm{VP}_c) - s(\mathrm{VP}_r)\big)\right)\right], \tag{4}$$

where $s(\mathrm{VP}_c)$ is the scalar output of the ranking network for the chosen visual prompt out of the pair, $s(\mathrm{VP}_r)$ represents the rejected one, and σ denotes the Sigmoid function.

3.4 Robust Inference Pipeline

During the inference, we input multiple candidate visual prompts along with the query into the large vision-language model and pass them through our ranking network, as shown in Figure 2. The network outputs a reward scalar for each visual prompt, and we select the highest-scoring one. Its visual tokens are concatenated with the text embeddings and decoded by the LLM to generate

the final answer. Besides, **to reduce distributional bias**, we introduce a pre-filtering step: the prompt whose visual features are farthest from other candidates, as measured by cosine distance, is removed in advance, further mitigating the negative effects of suboptimal prompts and enhancing robustness.

4 Experiments

4.1 Experimental Setup

Model architectures. We apply AutoV to four open-source LVLMs, including the LLaVA-1.5 series [32], LLaVA-OneVision [26], Qwen2.5-VL [5], and InternVL2 [14]. In particular, the LLaVA-1.5 builds on Llama2 [48]. LLaVA-OneVision enhances it with SigLIP [61] and Qwen2 [3]. Qwen2.5-VL introduces a dynamic-resolution encoder, and InternVL2 integrates InternViT [15] with InternLM2 [11].
Evaluation benchmarks. AutoV is extensively evaluated on **fourteen** vision-language benchmarks, covering video reasoning (CVBench [45]), vision-centric tests (BLINK [19], MMVP [46], MMStar [13]), comprehensive multimodal evaluation (MM-Vet [59]), knowledge/math reasoning (MMMU [60], MathVista [33]), robustness and VQA (VizWiz [9], Real-World QA [52], TextVQA [41], LLaVA-Wild [31]), grounding (RefCOCO+ [56]), captioning (COCO-Caption2017 [29]), and multimodal classification (ImageNet-1K [16]).

4.2 Main Results

Results are presented in Table 1, and we compare AutoV with three prior visual prompting methods for LVLMs, namely FGVP [55], RedCircle [40], and API [57]. The training data and implementation details can be found in Appendix A and C, respectively. Here, we employ the above three types of visual prompts as the candidate pool[4] for AutoV to retrieve from optimally.

The four main observations are as follows: (1) For all models, the average improvement of AutoV relative to the base models is 3.5%. Our method performs particularly well for LLaVA-OneVision on VizWiz, with an improvement of **10.2**%. (2) AutoV outperforms existing visual prompt engineering methods. For instance, on MMMU, it achieves an average gain of **3.2**%, compared to **1.1**% for API, yielding a **2.1**% absolute improvement. This margin highlights the importance of accurate case-level visual cue retrieval, as opposed to benchmark-level prompt design. In contrast, other visual prompting strategies that lack query-adaptive mechanisms tend to underperform. (3) Our method further supplements inherently less accurate models. For instance, LLaVA-7B with AutoV achieves higher accuracy than the 13B-scale model on **4 out of 7 tasks**. This means that AutoV can be utilized for model efficiency on certain specific tasks: **a few additional MLP layers are a better choice compared with hundreds of times larger parameter sizes.** (4) Both Qwen2.5-VL and InternVL2

[4] Includes four prompts extracted from different transformer layers of CLIP in API [57], with two additional prompts of distinct types, resulting in six images in total.

Table 1: Comparison of AutoV with other visual prompts methods for various LVLMs. Red indicates a decrease, while green represents an improvement.

Method	MMMU	VizWiz	MMVet	VQA^{Text}	LLaVA^{Wild}	MMStar	MathVista
			🤗 **(LLaVA-1.5 7B)**				
Base	36.3	50.0	30.6	58.2	65.4	33.2	34.2
FGVP	36.9 +0.6	47.5 −2.5	29.0 −1.6	46.3 −11.9	55.6 −9.8	33.0 −0.2	34.6 +0.4
Circle	37.0 +0.7	49.6 −0.4	31.3 +0.7	55.5 −2.7	62.2 −3.2	33.4 +0.2	35.0 +0.8
API	37.4 +1.1	51.3 +1.3	31.8 +1.2	58.4 +0.2	67.2 +1.8	34.0 +0.8	35.7 +1.5
AutoV	**38.7** +2.4	**52.1** +2.1	**32.9** +2.3	**60.6** +2.4	**68.6** +3.2	**35.2** +2.0	**36.9** +2.7
			🤗 **(LLaVA-1.5 13B)**				
Base	36.7	53.6	36.1	61.2	66.6	32.8	35.0
FGVP	36.9 +0.2	46.8 −6.8	29.7 −6.4	49.2 −12.0	63.7 −2.9	31.9 −0.9	35.3 +0.3
Circle	36.9 +0.2	53.8 +0.2	31.2 −4.9	58.1 −3.1	66.9 +0.3	33.5 +0.7	36.2 +1.2
API	37.3 +0.6	54.6 +1.0	37.6 +1.5	62.1 +0.9	67.6 +1.0	33.0 +0.2	37.0 +2.0
AutoV	**38.9** +2.2	**56.3** +2.7	**38.0** +1.9	**62.7** +1.5	**69.5** +2.9	**34.2** +1.4	**38.3** +3.3
			🤗 **(LLaVA-OneVision 7B)**				
Base	49.4	58.2	48.8	76.1	80.6	61.8	63.2
FGVP	49.6 +0.2	58.9 +0.7	49.1 +0.3	65.4 −10.7	70.8 −9.8	58.8 −3.0	63.4 +0.2
Circle	50.4 +1.0	60.2 +2.0	47.2 +1.6	77.2 +1.1	80.8 +0.2	62.7 +0.9	65.4 +2.2
API	50.8 +1.4	66.9 +8.7	51.5 +2.7	77.7 +1.6	81.9 +1.3	63.0 +1.2	64.9 +1.7
AutoV	**54.0** +4.6	**68.4** +10.2	**52.9** +4.1	**78.0** +1.9	**85.3** +4.7	**64.5** +2.7	**66.8** +3.6
			(InternVL2 8B)				
Base	48.6	58.6	46.5	77.0	74.6	62.0	58.3
FGVP	49.6 +1.0	58.3 −0.3	44.3 −2.2	70.7 −6.3	75.7 +1.1	60.0 −2.0	58.9 +0.6
Circle	48.3 −0.3	57.5 −1.1	45.8 −0.7	74.9 −2.1	75.4 +0.8	63.4 +1.4	58.4 +0.1
API	50.5 +1.9	61.2 +2.6	48.6 +2.1	77.0 +0.0	76.2 +1.6	63.9 +1.9	59.4 +1.1
AutoV	**51.7** +3.1	**64.8** +6.2	**50.2** +3.7	**78.0** +1.0	**79.5** +4.9	**65.8** +3.8	**70.5** +2.2
			(Qwen2.5-VL 7B)				
Base	50.1	69.5	56.6	83.0	98.0	62.4	68.2
FGVP	49.6 −0.5	70.8 +1.3	43.1 −13.5	78.0 −5.0	98.2 +0.2	63.4 +1.0	68.0 −0.2
Circle	51.6 +1.5	69.2 −0.3	44.5 −12.1	80.5 −2.5	97.5 −0.5	65.1 +2.7	69.1 +0.9
API	51.0 +0.9	72.1 +2.6	58.0 +1.4	85.3 +2.3	100.6 +2.6	63.0 +0.6	67.6 −0.6
AutoV	**53.9** +3.8	**74.4** +4.9	**59.1** +2.5	**86.1** +3.1	**102.3** +4.3	**65.6** +3.2	**70.2** +2.0

directly adopt the retrieval optimal prompting strategy from AutoV, which is pre-trained on LLaVA-OneVision. This integration results in an average accuracy gain of **4.2**%. This demonstrates that the retrieval-enhanced prompting strategy is **model-agnostic** and generalizes well across heterogeneous LVLM architectures, regardless of pretraining scales or vision encoders. Additionally, we extend the evaluation to a broader range of vision-language benchmarks in Appendix E, highlighting AutoV's strong cross-task generalization in retrieval.

5 Analysis

5.1 Comparison with Other Retrieval Methods

To analyze why pair-wise ranking is more suitable for visual prompt retrieval, we compare it with several representative retrieval strategies. All experiments are conducted on LLaVA-7B with the retrieval pool utilized in the main experiments.

Table 2: Analysis between AutoV and other retrieval methods. VP_1-VP_4 are from attention prompting [57] from different layers.

Method	MMMU	VizWiz	MMVet	Avg.
Baseline	36.3	50.0	30.6	39.0
VP_1 only	36.7 +0.4	50.7 +0.7	31.8 +1.2	39.8 +0.8
VP_2 only	36.9 +0.6	51.3 +1.3	31.4 +0.8	39.9 +0.9
VP_3 only	37.4 +1.1	50.6 +0.6	31.7 +1.1	39.9 +0.9
VP_4 only	37.2 +0.9	51.2 +1.2	31.2 +0.6	39.9 +0.9
Random	37.1 +0.8	50.7 +0.7	31.3 +0.7	39.7 +0.7
Regression	36.9 +0.6	50.9 +0.9	31.1 +0.5	39.6 +0.6
MoE	37.6 +1.3	50.7 +0.7	32.0 +1.4	40.1 +1.1
AutoV List-wise ranking	38.0 +1.7	51.3 +1.3	32.4 +1.8	40.6 +1.6
AutoV Pair-wise ranking	**38.7** +2.4	**52.1** +2.1	**32.9** +2.3	**41.3** +2.3

1. Heuristic Fixed Retrieval. This strategy selects a single predefined visual prompt based on manually designed rules. As shown in Rows 3–6 of Table 2, different prompts exhibit distinct performance profiles across benchmarks. For instance, VP_3 performs best on MMMU, while VP_1 is optimal on MMVet, indicating that no single prompt engineering is universally effective.

2. Random Retrieval. Randomly selecting yields a +0.7% average gain over the baseline and performs on par with or slightly better than individual fixed prompts. This suggests that prompt retrieval itself provides benefits.

3. Regression Retrieval. This strategy directly predicts the absolute loss score of each visual prompt and selects the lowest one. As shown in Table 2, regression improves the baseline by 0.6%, but performs worse than both MoE and ranking-based methods. This suggests that learning absolute loss is challenging due to small performance gaps among prompts and inherent noise in supervision signals, making it difficult to reliably distinguish the optimal prompt.

4. MoE (GateNet) Retrieval. We further compare with a Mixture-of-Experts (MoE) style retrieval mechanism, where a lightweight gating network (implemented with Gumbel Softmax [18]) learns to select among multiple visual prompts in a differentiable manner. GateNet is inserted after the projection layer and trained under the same data and training budget as AutoV. As shown in Row 9 of Table 2, MoE improves the baseline by 1.1%. While this demonstrates the benefit of learnable selection, its gain over random retrieval is modest (+0.4%), and it remains 1.2% behind AutoV (pair-wise ranking).

5. List-wise Ranking. We further adopt a list-wise ranking objective that models the relative ordering of all candidate prompts simultaneously. List-wise ranking achieves consistent gains (+1.6%), demonstrating the benefit of structured preference learning over regression or soft gating. However, compared to pair-wise ranking, its improvement is more limited. We conjecture that list-wise optimization requires estimating a full permutation over candidates, which introduces higher optimization complexity and sensitivity to noisy comparisons.

Table 3: Performance of various sizes of prompts pool. 1–4 are attention prompting images [57] from different layers; 5 and 6 come from [40] and [55]; 7 and 8 are **newly unseen prompts** ([24] and [65]) for analyzing scaling law effects.

Pool Size	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Source	+API-L_{15}	+API-L_{20}	+API-L_{22}	+API-L_{23}	+RedCircle	+FGVP	+VTP	+TVP
MMMU	$37.3_{+1.0}$	$37.5_{+1.2}$	$38.3_{+2.0}$	$38.3_{+2.0}$	$38.6_{+2.3}$	$38.7_{+2.4}$	$38.7_{+2.4}$	$\mathbf{38.9}_{+2.6}$
VizWiz	$50.3_{+0.3}$	$50.8_{+0.8}$	$51.0_{+1.0}$	$51.9_{+1.9}$	$52.1_{+2.1}$	$52.1_{+2.1}$	$52.6_{+2.6}$	$\mathbf{52.6}_{+2.6}$
MMVet	$31.8_{+1.2}$	$32.0_{+1.4}$	$32.6_{+2.0}$	$32.8_{+2.2}$	$32.9_{+2.3}$	$32.9_{+2.3}$	$33.1_{+2.5}$	$\mathbf{33.1}_{+2.5}$
Avg.	$39.8_{+0.8}$	$40.1_{+1.1}$	$40.7_{+1.7}$	$41.0_{+2.0}$	$41.2_{+2.2}$	$41.3_{+2.3}$	$41.5_{+2.5}$	$\mathbf{41.6}_{+2.6}$

6. AutoV. Overall, these comparisons suggest that effective visual prompt retrieval requires (i) query-aware adaptability beyond random diversity, (ii) explicit relative preference modeling beyond soft gating, and (iii) structured supervision beyond list-wise ranking. Pair-wise ranking naturally satisfies these properties, which explains its consistent superiority in Table 2.

5.2 Retrieval Visualization

We visualize the prompt distribution of retrieval strategies within VP_1-VP_4 for significance. Figure 4 presents a bar graph comparing the retrieval decisions of GateNet and our AutoV on MMVet. The bar heights indicate the frequency of visual prompt retrieval, while the overlaid line plots depict the independent accuracy for each prompt. Our analysis reveals a notable discrepancy between the GateNet retrieval and the actual inference accuracy of visual prompts. For instance, GateNet retrieves VP_3, the second-most useful prompt, only 38 times (38/218, 13.6% of cases), despite its high accuracy. In contrast, AutoV exhibits stronger alignment with the standalone accuracy of visual prompts, and prefers to select VP_1 and VP_3, the two prompts that own superior accuracy.

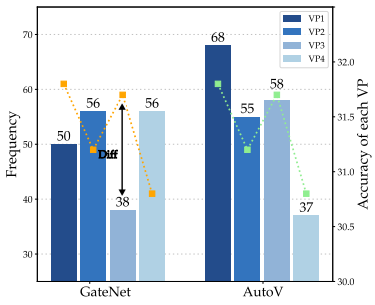


Fig. 4: Retrieval distribution on MMVet. The plots are independent accuracy for each prompt.

5.3 The Scaling Law of AutoV

Having more visual prompt candidates increases selection diversity and potentially improves the adaptability of LVLMs. **However, is more always better?** Here, we extend the candidate pool from 1 to 8 prompts (Table 3). **(1)** We observe a clear diminishing-return effect. When prompts are individually beneficial, increasing their number initially improves accuracy, but the marginal gain decreases as the pool grows. For example, on MMMU, the gain remains 0.8%

Table 4: Ablation studies on LLM layer selection and pre-filtering step.

Method	MMMU	VizWiz	MMVet
LLaVA-1.5-7B	36.3	50.0	30.6
Part1: LLM layer selection			
24th layer	38.1	51.6	32.3
12th layer	38.4	51.7	32.3
0th layer (**Default**)	**38.7**	**52.1**	**32.9**
Part2: Pre-filtering strategy			
When 6 Candidates			
w/o pre-filtering	38.3	52.0	32.6
w/ pre-filtering (**Default**)	**38.7**$_{+0.4}$	**52.1**$_{+0.1}$	**32.9**$_{+0.3}$
When 8 Candidates			
w/o pre-filtering	38.4	52.2	32.7
w/ pre-filtering (**Default**)	**38.9**$_{+0.5}$	**52.6**$_{+0.4}$	**33.1**$_{+0.4}$

when increasing from 2 to 3 prompts, but drops to 0 when expanding from 6 to 7 prompts. **(2)** AutoV is largely agnostic to the quality of individual visual prompts, and its effectiveness does not depend on carefully curated prompt candidates. On VizWiz, although RedCircle and FGVP yield negative gains (-0.4% and -2.5%), incorporating them still leads to a 0.2% improvement, demonstrating its robustness. **(3)** Moreover, even when encountering unseen prompt types during training, AutoV can effectively index suitable candidates, yielding an additional 0.3% average improvement on the last two visual prompts.

Overall, AutoV remains stable under varying prompt quality and pool sizes, consistently preserving or improving performance. In practice, we recommend using high-quality prompts while limiting the pool to fewer than 9 candidates to balance effectiveness and efficiency.

5.4 Study on the LLM Layer Selection

Findings from FastV [12] and LLaVA-Mini [64] validate that cross-modal interactions mainly occur in shallow layers. This enables the textual input to better guide visual selection. **Part 1** of Table 4 shows that the 0th layer outperforms intermediate or bottom layers, validating our design choice.

5.5 Study on the Pre-filtering Step

Part 2 of Table 4 shows consistent improvements, especially with more prompts. High-recall pre-filtering removes only the most dissimilar candidates, with a negligible risk of discarding optimal prompts ($<0.5\%$ in our statistics), while substantially reducing the presence of harmful prompts.

5.6 The Transferability of AutoV

The retrieval strategy in our AutoV, trained on one LVLM, can be transferred to other LVLMs. First, our method demonstrates broad applicability across open-

Table 5: Performance of closed-source models with AutoV. The retrieval strategy is generated from AutoV pre-trained on LLaVA-OneVision.

Models	VizWiz	$\text{LLaVA}^{\text{Wild}}$
Gemini (Gemini-1.5-Pro)	45.9	77.9
+AutoV	$55.5_{+9.6}$	$81.1_{+3.2}$
ChatGPT (GPT-4o)	52.8	68.0
+AutoV	$61.8_{+9.0}$	$70.9_{+2.9}$
Qwen-VL (Qwen-VL-Max)	64.6	85.8
+AutoV	$71.9_{+7.3}$	$88.4_{+2.6}$

Table 6: Overhead analysis. Extra and total FLOPs (T) under different pool sizes, model scales, and resolutions.

Factor	Base Setting		Expanded Setting	
	Extra (T)	**Total** (T)	**Extra** (T)	**Total** (T)
Pool Size	0.74 (**4**)	5.08	1.62 (**8** ↑)	5.96
Model Size	0.74 (**7B**)	5.08	0.74 (**13B** ↑)	8.77
Vision Token	0.74 (**576**)	5.08	1.48 (**1152** ↑)	6.59

source models. In Table 1, both InternVL2 and Qwen2.5-VL achieve a 3.2%+ improvement by directly adopting the retrieval strategy from AutoV pre-trained on LLaVA-OneVision. Besides, Table 5 reveals that even closed-source models like Gemini-Pro [44] and GPT-4o [23] show substantial gains, averaging 6.4 and 6.0 points respectively. Therefore, AutoV owns good transferability, as its retrieval strategy consistently boosts performance across both open- and closed-source LVLMs without model-specific adaptation.

5.7 Statistical Significance Analysis

To evaluate the reliability of our improvements, we perform paired t-tests [21] across multiple random seeds for all benchmarks in Table 1 (LLaVA-1.5 7B). Taking MMMU as an example, AutoV achieves a mean improvement of $\Delta = +2.06\%$ over the baseline with $p \leq 0.05$, indicating statistical significance rather than random fluctuation. Similar significance is consistently observed across other benchmarks. Detailed statistical results are provided in Appendix G.

5.8 The Efficiency of AutoV

AutoV improves LVLM performance while remaining lightweight and practical. **Training** AutoV on LLaVA-7B takes only 6 hours on 8 × A100GB GPUs (40 epochs). Thanks to its transferability, no retraining is required for new LVLMs, and training data can be generated with smaller models. **At inference**, AutoV encodes $N-1$ additional images and ranks N candidates ($N \leq 8$), without extra decoder cost. For LLaVA-7B (576 visual, 20 text tokens), the encoder, ranking-net, and decoder require 0.16T, 0.09T, and 4.30T FLOPs, respectively:

Fig. 5: Visualization of retrieval results faced with different queries. VP_1 and VP_3 come from API, VP_2 is from RedCircle, and VP_4 is from FGVP.

LLM decoding clearly dominates. Under the base setting ($N = 4$), AutoV adds only 0.74T FLOPs (Table 6). Larger pools or higher resolution increase vision-side cost nearly linearly, yet remain negligible relative to decoder computation. The overhead is independent of model scale (7B vs. 13B). In practice, **latency** increases by just 6.9 ms (254.8 ms $\rightarrow$ 261.7 ms) due to compact similarity computation and early prompt selection, yielding a $+2.4\%$ average accuracy gain.

5.9 Qualitative Visualization

To validate the effectiveness of our retrieval strategy, we visualize an example sourced from LLaVA-Bench within {API-L22-Dark, RedCircle, API-L23-White, FGVP}. As shown in Figure 5, we list the output results of AutoV when we enter different queries. When faced with a question that focuses on the global context information, AutoV chooses VP_1 (API-L22-Dark), where the black mask removes the background. While referring to the "intended effect" in the image, AutoV selects VP_2 (RedCircle), which directly uses red circles to highlight the weird spots. More visualizations can be found in Appendix H.

6 Conclusion

In this paper, we introduce AutoV, an adaptive framework that retrieves optimal visual prompts for textual queries in LVLMs. Leveraging a lightweight ranking network trained with reward-based supervision, AutoV consistently outperforms prompt engineering methods. Notably, its effectiveness does not rely on the intrinsic quality of individual visual prompts, nor is it constrained by their standalone performance. Extensive experiments on multiple LVLMs and benchmarks demonstrate its effectiveness, robustness, and generality.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62476011 and 62472008, and by the Beijing Natural Science Foundation under Grant L252060.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv:2303.08774 (2023) 1, 3
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems (2022) 3
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv:2309.16609 (2023) 1, 8
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A frontier large vision-language model with versatile abilities. arXiv:2308.12966 (2023) 1, 3
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 1, 3, 8
6. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024) 3
7. Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., Wei, F.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems (2022) 4
8. Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al.: Deepseek LLM: Scaling open-source language models with longtermism. arXiv:2401.02954 (2024) 1, 3
9. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., et al.: Vizwiz: nearly real-time answers to visual questions. In: ACM symposium on User interface software and technology (2010) 8
10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems (2020) 1, 3
11. Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., et al.: Internlm2 technical report. arXiv preprint arXiv:2403.17297 (2024) 3, 8
12. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision (2024) 12
13. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems (2024) 8
14. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024) 8
15. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2024) 1, 3, 8

16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2009) 8
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020) 3
18. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research (2022) 4, 10
19. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision (2024) 8
20. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (2017) 6
21. Hsu, H., Lachenbruch, P.A.: Paired t test. Wiley StatsRef: statistics reference online (2014) 13
22. Huang, W., Liu, H., Guo, M., Gong, N.: Visual hallucinations of multi-modal large language models. In: Findings of the Association for Computational Linguistics (2024) 3
23. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 1, 13
24. Jiang, S., Zhang, Y., Zhou, C., Jin, Y., Feng, Y., Wu, J., Liu, Z.: Joint visual and text prompting for improved object-centric perception with multimodal large language models. arXiv preprint arXiv:2404.04514 (2024) 3, 11
25. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems (2022) 2
26. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2024) 1, 3, 8
27. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (2023) 3
28. Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., Jia, J.: Mini-gemini: Mining the potential of multi-modality vision language models. arXiv:2403.18814 (2024) 3
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014) 8
30. Lin, W., Wei, X., An, R., Peng, G., Zou, B., Luo, Y., Huang, S., Zhang, S., Li, H.: Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. In: International Conference on Learning Representations (2025) 2, 3, 4
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2024) 1, 3, 8
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems (2024) 8

33. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: International Conference on Learning Representations (2024) 8
34. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (2022) 7
35. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv:2304.03277 (2023) 1, 3
36. Qi, J., Xu, Z., Shao, R., Chen, Y., Di, J., Cheng, Y., Wang, Q., Huang, L.: Rora-vlm: Robust retrieval-augmented vision language models. arXiv preprint arXiv:2410.08876 (2024) 4
37. Qu, L., Li, H., Wang, T., Wang, W., Li, Y., Nie, L., Chua, T.S.: Unified text-to-image generation and retrieval. arXiv preprint arXiv:2406.05814 (2024) 3
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021) 4, 5
39. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog (2019) 3
40. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2023) 3, 4, 8, 11
41. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards VQA models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 8
42. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-clip: A clip model focusing on wherever you want. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2024) 2, 3, 4
43. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv:2312.11805 (2023) 1
44. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024) 1, 13
45. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems (2024) 8
46. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2024) 8
47. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv:2302.13971 (2023) 1, 3
48. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 8

49. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems (2022) 2
50. Wu, J., Li, X., Yu, T., Wang, Y., Chen, X., Gu, J., Yao, L., Shang, J., McAuley, J.: Commit: Coordinated instruction tuning for multimodal large language models. arXiv preprint arXiv:2407.20454 (2024) 3
51. Wu, J., Zhang, Z., Xia, Y., Li, X., Xia, Z., Chang, A., Yu, T., Kim, S., Rossi, R.A., Zhang, R., et al.: Visual prompting in multimodal large language models: A survey. arXiv preprint arXiv:2409.15310 (2024) 2, 4
52. xAI: RealWorldQA. https://huggingface.co/datasets/xai-org/RealworldQA (2024), accessed: 2025-04-26 8
53. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 13204–13214 (2024) 6
54. Xing, S., Li, P., Wang, Y., Bai, R., Wang, Y., Hu, C.W., Qian, C., Yao, H., Tu, Z.: Re-align: Aligning vision language models via retrieval-augmented direct preference optimization. In: Empirical Methods in Natural Language Processing (2025) 4
55. Yang, L., Wang, Y., Li, X., Wang, X., Yang, J.: Fine-grained visual prompting. Advances in Neural Information Processing Systems (2023) 3, 4, 8, 11
56. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision (2016) 8
57. Yu, R., Yu, W., Wang, X.: Attention prompting on image for large vision-language models. In: European Conference on Computer Vision (2024) 3, 4, 8, 10, 11
58. Yu, S., Jin, C., Wang, H., Chen, Z., Jin, S., Zuo, Z., Xu, X., Sun, Z., Zhang, B., Wu, J., et al.: Frame-voyager: Learning to query frames for video large language models. In: International Conference on Learning Representations (2025) 6
59. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. In: International Conference on Machine Learning (2024) 8
60. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2024) 8
61. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2023) 8
62. Zhang, A., Fei, H., Yao, Y., Ji, W., Li, L., Liu, Z., Chua, T.S.: Vpgtrans: Transfer visual prompt generator across llms. Advances in Neural Information Processing Systems (2023) 2, 4
63. Zhang, C., Li, W., Ouyang, W., Wang, Q., Kim, W.S., Hong, S.: Referring expression comprehension with semantic visual relationship and word mapping. In: ACM International Conference on Multimedia (2019) 4
64. Zhang, S., Fang, Q., Yang, Y., Feng, Y.: Llava-mini: Efficient image and video large multimodal models with one vision token. In: International Conference on Learning Representations (2025) 5, 12
65. Zhang, Y., Dong, Y., Zhang, S., Min, T., Su, H., Zhu, J.: Exploring the transferability of visual prompting for multimodal large language models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2024) 3, 11

66. Zhang, Y., Huang, T., Fan, C.K., Dong, H., Li, J., Wang, J., Cheng, K., Zhang, S., Guo, H., et al.: Unveiling the tapestry of consistency in large vision-language models. Advances in Neural Information Processing Systems (2024) 6
67. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2022) 2
68. Zong, Z., Ma, B., Shen, D., Song, G., Shao, H., Jiang, D., Li, H., Liu, Y.: Mova: Adapting mixture of vision experts to multimodal context. Advances in Neural Information Processing Systems (2024) 4