

# Vision Language Models Learn to Assess Images with Specialists

Quy<sup>1</sup>, Seunghyun Yoon<sup>2</sup>, Ruiyi Zhang<sup>2</sup>, Thiloshon Nagarajah<sup>2</sup>, Trung Bui<sup>2</sup>, Viet Dac Lai<sup>2</sup>

<sup>1</sup> Virginia Tech    <sup>2</sup> Adobe Research

quyetdo@vt.edu, daclai@adobe.com

## Abstract

Automatic preference models play an important role in the development of image generation and image editing models, yet existing approaches often lack either generalizability or proficiency. To cope with this limitation, we argue that incorporating judgments from image assessment specialists can equip a VLM-as-a-judge with richer, more reliable understanding of images, ultimately enabling better modeling of human preference while preserving broad generalizability. With the additional specialist assessments, a VLM judge reasons through the assessments, identifies the key deciding factors, and generates the final preference judgment. We extensively experimented on four public benchmarks, covering both image generation and image editing tasks in both pointwise and pairwise preference paradigms. Our specialist-aided in-context learning (ICL) models improve alignment with human preferences by 2% to 8% over their corresponding baseline VLMs. Beyond ICL, we also investigate how specialists can help generate useful image preference data for VLMs. We reverse-engineer chain-of-thoughts image preference data, given the input, ground-truth preference, and specialists’ assessments; then train a VLM model on the synthetic data. Our finetuning model boosts the alignment by a margin of up to 13% over its corresponding baseline VLMs, achieving the best performance on certain benchmarks. Notably, it achieves a better overall performance than a State-of-The-Art image preference model with much less finetuning data. Overall, our work extensively shows the potential of combining VLMs with image assessment specialists for reliable image preference modeling.

## 1. Introduction

Generalizable and reliable preference models that align closely with human preferences are always desired for faster development of generative models [2, 20]. In text-to-image generation and text-guided image editing, VLM-based met-

\*Work done during the internship at Adobe Research

**Generation Request:** overgrown city street atmospheric, kodak, fuji film, photoreal, 12k ursa, volumetric light, cinematic photograph concept art, intricate, artstation, studio ghibli, eddie mendoza, james chadderton

**Human Score (continuous, 0 to 5):** 3.02

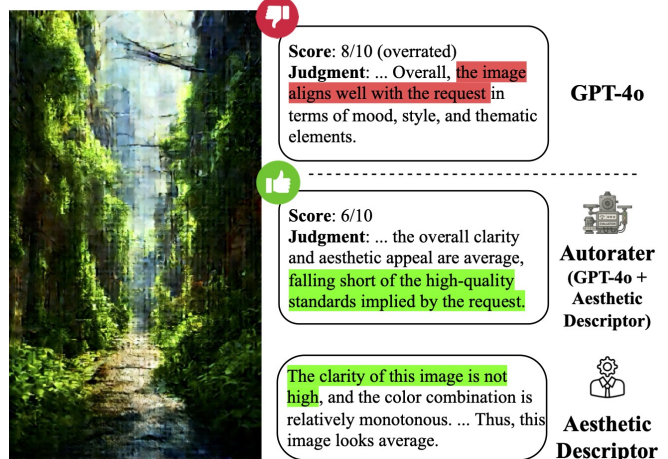


Figure 1. Evaluating an image generation sample using GPT-4o and our system, so-called AUTORATER, which has an aesthetic description tool to aid GPT-4o’s context. As the details in the image are blurred and not photorealistic, the human score of 3.02 out of 5 is reasonable. GPT-4o overrates the image as good-quality enough to satisfy the generation request; meanwhile, AUTORATER offers a better preference score with the support from an aesthetic descriptor.

rics (CLIP-Score [6], BLIP-Score [5]), and specialists on image preference<sup>1</sup> (PickScore [13], ImageReward [30], Q-SiT [35], EditReward [28]) have been widely adopted as proxies for human preference modeling. Despite their success in various scenarios, these models are still prone to deficiencies in out-of-distribution cases. For example, Q-SiT fails to align with human preference in the ImageRe-

<sup>1</sup>We use the word “preference” to indicate the scoring or the selection of a better one among a few images. Meanwhile, “assessment” means general assessment, which is a hypernym of “preference”.

ward benchmark [30] with a Pearson correlation of 0.34, while it achieves a state-of-the-art Pearson correlation of 0.85 in a similar benchmark AIGQA-3K [15]. Meanwhile, using general-purpose multi-modal large language models (VLMs) as preference models offers greater generalizability. However, they are not specialized for fine-grained image preference assessments and may overlook important visual details such as aesthetic [22] and identity preservation [11], as shown in Figure 1. This presents a gap in developing highly generalized, accurate image preference assessment models.

In this paper, we aim to address the aforementioned gap by exploring the combination of VLMs and image assessment specialists as a unified preference model. To this end, we harness the broad generalization and reasoning capabilities of VLMs while benefiting from the high-fidelity, domain-specific evaluation strengths offered by specialist models. In our framework, we not only consider image preference specialists that directly predict overall preference, but also incorporate specialists that assess images along key dimensions such as text-image alignment, image quality, and fine-detail preservation.

In particular, we present AUTORATER, a framework that integrates specialists to assess images along key dimensions, i.e., text-image alignment, image quality, and detail preservation. Subsequently, a general VLM judge reasons over the input and these assessments to produce the final preference. The final preference includes a natural-language analysis and either a score or a comparison, corresponding to either a pointwise or a pairwise preference. In this work, we analyze three specialist selection mechanisms: 1) All specialists that are available for the case; 2) Automatic selection for each sample; and 3) Grid-Search for an optimal subset of specialists. Beyond the in-context learning (ICL) method, we also explore whether VLMs can learn from specialists to become better preference models. We synthesize chain-of-thought preference assessment data from the input, ground-truth preferences, and specialist assessment; then finetune a VLM on this synthetic dataset.

We validate the variants of AUTORATER and baselines versus human preference on three benchmarks [15, 25, 30], covering both image generation and editing datasets and both pointwise and pairwise preference paradigms. ICL variants of AUTORATER improve the alignment with human preferences by 2% to 8% over its corresponding baseline VLMs in all benchmarks. Meanwhile, the finetuning approach further boosts the alignment of AUTORATER by a large margin up to 13% over its corresponding baseline VLMs, achieving the best performance on ImageReward [30] and AIGQA-30K [15] in pointwise preference setting. In overall, it achieves a better overall performance than a State-of-The-Art image preference model HPSv3

[17], while *using much less finetuning data*. Furthermore, both ICL and finetuning approaches perform better than Q-SiT, a strong baseline yet also an image assessment specialist for AUTORATER, by a large margin. That means combinations of VLMs and specialists can exceed the capability of each individual model. Clearly, AUTORATER sustains competitiveness across the editing and generation tasks and pointwise and pairwise settings, while Q-SiT failed to generalize on new data distribution and new evaluation scenarios. Our qualitative analysis shows examples where specialists contribute accurate low-level details that help VLMs produce better preference assessments. It also shows examples where such contributions are ineffective or introduce noise that harms performance.

In summary, we make the following key contributions:

- We proposed a *specialist-aided VLM-as-a-judge framework for image preference* assessment that leads to better image preference models that align closer with human preference while maintaining strong generalizability on new preference tasks.
- Our detailed analysis shows that specialists provide valuable information for the base VLM in identifying the deciding factor in human preference decisions.
- We showed that synthetic specialist-guided rationales provide a substantial reasoning signal for *efficiently finetuning* a preference model, leading to the best models for pointwise preference tasks while maintaining relatively good performance on pairwise preference tasks.

## 2. Related Works

**Text-to-image generation:** Pick-A-Pic [13] and HPD v2 [29] provide human preference data as the most preferred one among a group or a pair of generated images. Meanwhile, MHP [33] and ImageRewardDB [30] contain human annotations, including overall rating and rating for each assessment dimension of generated images. More recently, AIGQA-30K [15] focuses on the mean opinion score (MOS) of generated images regarding perceptual quality and alignment. Q-Bench<sup>+</sup> [34], on the other hand, proposes an evaluation suite on low-level details for multi-modal foundation models.

**Text-guided image editing:** EmuEdit [24] leverages LLMs to generate editing instructions and uses Stable Diffusion models to synthesize images. On the other hand, MagicBrush [31] has annotators propose edits for given images and uses the DALL-E 2 platform to iteratively make the edits. Recently, PSR [25], a dataset collected from the subreddit Photoshop Request, was released. A subset of PSR has human annotation on the better edited image, which can serve as ground truth to verify the automatic evaluation pipeline. Those datasets that contain human preferences for the output images serve as benchmarks to measure alignments between image preference models with human pref-

erences. *In this paper, we consider AIGQA-30K, ImageReward, and PSR in our experiments.*

**Image Assessment Models:** Earlier research has utilized human annotations, VLM-based judging systems, or image assessment models for preference assessment purposes. With the popularity of text-to-image generation, many vision-language pre-training models have been adopted to evaluate the text-image alignment, including: CLIP-Score [6], BLIP-Score [5], HPSv2 [29], PickScore [13], ImageReward [30]. These models typically assess overall preference for text-to-image generations by considering multiple contributing factors. Apart from that, these models are also adopted by many image editing benchmarks.

Extended from previous works, MPS [33] introduces a pretrained preference model on multiple dimensions, including aesthetics, details, and alignments. Similarly, HPSv3 [17] leverages a wide-spectrum million-sample human preference dataset to train preference models. MINT-IQA [26], on the other hand, approaches the image preference problem via instruction tuning. Meanwhile, Q-SiT [35] is jointly trained for image quality scoring and interpretation. However, they are all vulnerable to the classic generalizability problem of pretrained AI specialists. Our results show that Q-SiT, as a leading model for preference scoring, drastically fails in new evaluation settings.

With the advancement of VLMs, such as GPT-4 [18], a line of work leverages VLMs’ generalizability to design new image assessment metrics for various evaluation paradigms. Those includes VIEScore [14], HQ-Edit [10], and T2I-CompBench++ [8]. *In this work, we explore the combination of VLMs with specialist models.* The goal is to leverage the generalizability of VLMs alongside the specialists’ proficiency in low-level detail assessment.

### 3. Problem Formulation

We consider both text-to-image generation and text-guided image editing tasks for evaluation. We denote a sample of task T as  $(I, O, a)$ , where:

- $I$  is a set of all necessary inputs for the task, including the textual request and a source image (if T is the text-guided image editing task),
- $O$  is a set of output images edited or generated by either a human, tools, or a generative model.
- $a$  is the ground-truth human preference.

We study automatic evaluation protocols  $A$  that take  $\{I, O\}$  as input and yield a preference assessment that aligns well with human annotation  $a$ . We denote the **objective** as:

$$A(I, O) \sim a.$$

We also denote a dataset of the task T as  $D = \{(I_i, O_i, a_i) | i = \overline{1, n}\}$ . The metric to validate our AUTORATER on  $D$  is described below for each case.

**Pairwise comparison:** In case there are **two output images**  $(O'_i, O''_i)$ ,  $a_i$  indicates which image is better. The alignment is represented by  $A$ ’s prediction accuracy in comparison to human annotations. The optimization objective is:

$$\arg \max_A \frac{1}{|D|} \sum_i \mathbb{I}(A(I_i, O_i) = a_i) \quad (1)$$

**Pointwise rating:** In case there is only **one output image**,  $a_i$  is a score for the output image. The alignment is then represented by the Pearson correlation  $r$  of  $A$ ’s outputs and human annotations. The optimization objective is:

$$\arg \max_A r(\{(A(I_i, O_i), a_i | i)\}) \quad (2)$$

Overall, **we consider three evaluation scenarios:** image generation with pointwise rating, image generation with pairwise comparison, and image editing with pairwise comparison. In the scope of this paper, we do not consider other evaluation paradigms, such as image generation with multiple instance ranking.

## 4. AUTORATER

Figure 2 outlines the overall design of our framework **AUTORATER**. In this section, we describe image assessment specialists and elaborate on three specialist selection mechanisms that we consider. Beyond ICL, we describe a fine-tuning pipeline that aims to enhance the chain-of-thoughts image preference assessment capability of VLM judges, given specialist assessments as extra context.

### 4.1. Image Assessment Specialists

In our system, we consider image assessment specialists which cover three common evaluation dimensions: text-image alignment [1, 7, 19], image quality [19, 21, 30], and detail preservation [12, 21]. Each specialist offers targeted capabilities to augment VLM judges. Below are descriptions of targeted capabilities and the considered tools.

#### Text-Image Alignment:

- **Object Detection:** *List of detected objects* can be used in determining the alignment of the output image with the generation/editing request. We use **GroundingDINOv2** [16], a powerful open-source object detection model, as the object detector. Because we want to focus on objects mentioned in the request, we pass the generation/editing request as the input caption for detection.

#### Image Quality:

- **Aesthetic Description:** Visual appeal and aesthetic details of generated or edited images are crucial for high-quality outputs. We use **AesExpert** [9], a model fine-tuned on large-scale aesthetic instruction data, to provide *nuanced aesthetic descriptions of images*. The descriptions cover aspects such as lighting, composition, color, and fidelity, which are useful for the VLM judge.

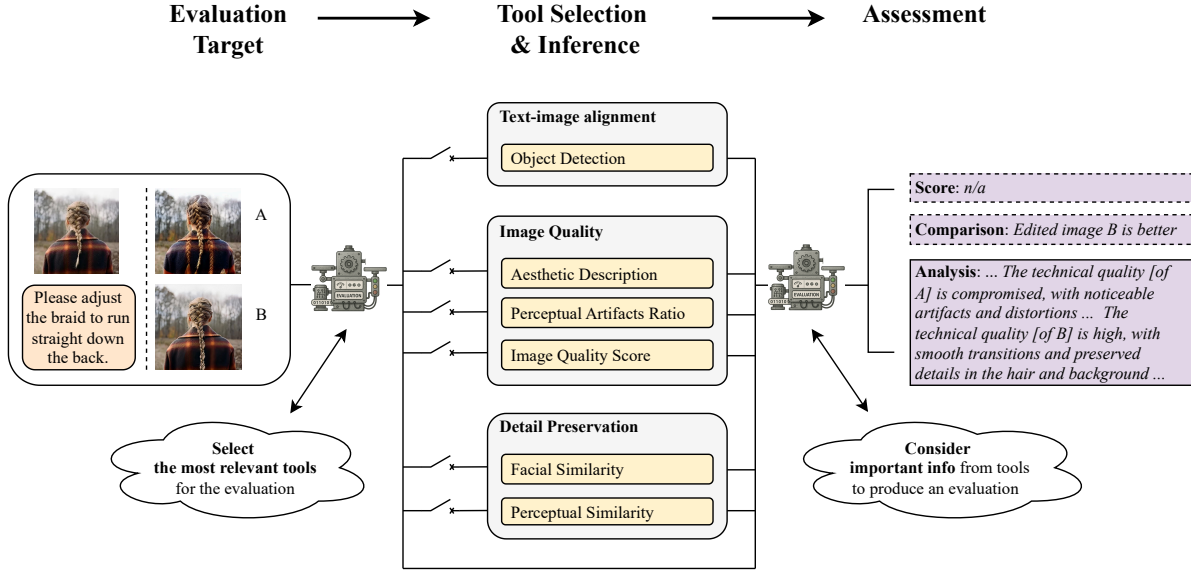


Figure 2. Overall architecture of AUTORATER, the system for our study. Given a generation or editing request, a source image in case of editing, and one or two generated or edited images as input, through a tool selection mechanism, extra information from a subset of specialists will be provided into the context. After that, AUTORATER uses a VLM judge to produce the final assessment, which includes a natural-language analysis and either a score or a comparison, depending on whether the evaluation paradigm is pointwise or pairwise.

Task	Type	Benchmark	Size	Metrics
Editing	Pair	PSR	444	Accuracy
	Pair	ImageReward	6,399	Accuracy
Generation	Point	ImageReward	2,720	Pearson
	Point	AIGQA-30K	4,000	Pearson

Table 1. Four benchmarks covering both image generation and editing tasks and both pointwise and pairwise evaluation paradigms.

- **Perceptual Artifacts Ratio:** Detecting perceptual artifacts such as distortions and unnatural blends is important to judge image fidelity. We employ **PAL4VST** [32], which localizes artifacts in generated or edited images, then compute the *artifacts ratio* (ranging from 0 to 1, with 0 indicating no artifacts). The score would aid the VLM judge in penalizing images with significant visual flaws.
- **Image Quality Score:** For a holistic assessment of technical quality, including sharpness and perceptual clarity, we use **Q-SiT** [35]. This model produces an **image quality score** (ranging from 0 to 1, where 1 means perfect), which provides a reference for our system to judge the overall image quality. It is a leading image assessment specialist that achieves SoTA performance on multiple image assessment benchmarks.

#### Detail Preservation (for image editing only):

- **Facial Similarity:** When an image editing task involves

human subjects, evaluating identity consistency is important. We leverage **DeepFace** [23], a comprehensive facial recognition and attribute analysis toolkit. It computes the **cosine similarity of faces** in the source image and edited image, representing facial similarity.

- **Perceptual Similarity:** In scenarios requiring minimal editing, perceptual coherence is important. We adopt **DreamSim** [4], which measures *perceptual similarity of the source image and edited image* beyond pixel-level metrics. It considers semantic structure, layout, and pose.

## 4.2. Specialist Selection Mechanisms for Evaluation

**All:** AUTORATER uses all tools that are available for each evaluation case. When the task is image generation, no matter whether the evaluation paradigm is pointwise or pairwise, we provide GroundingDINOv2 (GD), AesExpert (AE), PAL4VST (PA), and Q-SiT (QS) for AUTORATER. Similarly, for the image editing task, we provide the previously mentioned four tools for image generation, DeepFace (DF) and DreamSim (DS).

**Auto:** For each sample, among all tools that are available for the case, the VLM judge determines a subset of tools needed for the evaluation. Only assessments from selected tools will be included in the VLM judge’s context. We note that AUTORATER with this mechanism will require two VLM passes.

**Grid-Search:** For each benchmark and each base VLM, a subset of tools is selected for AUTORATER. To extensively explore the AUTORATER framework, we conduct

grid-search experiments. Every tool combination (single, subset of 2, subset of 3) will be evaluated. This tool selection mechanism is meant to explore maximal capability of our method in the ICL setup.

### 4.3. Learning to Judge

Beyond ICL, we investigate the capabilities of VLM-based judges when they are finetuned on *synthetic chain-of-thought reasoning data* generated for image preference assessment tasks. Each evaluation instance is augmented with all applicable specialist assessments that serve as additional context, enabling the model to learn rich, domain-informed rationales.

Extending the notation in section 3, let  $(S)$  denote the set of assessments produced by all applicable specialists for a given sample. For each training instance  $(I, O, a)$ , we prompt a strong base VLM, parameterized by  $\theta$ , to produce a synthetic rationale ( $\hat{r}$ ) explaining the ground-truth preference. The rationale is conditioned on the inputs (text, image), output images (if any), specialist feedback, and the ground-truth preference:

$$\hat{r} \sim \text{VLM}_\theta(I, O, S, a) \quad (3)$$

These generated rationales act as latent explanatory variables that guide model learning. To explicitly integrate generated rationale into finetuning, we cast the finetuning process as an expectation maximization optimization problem [3]. In the E-step, we obtain latent rationales ( $\hat{r}$ ); in the M-step, we update model parameters to maximize the likelihood of producing both the rationale and the ground-truth preference:

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{\hat{r}} \log P(\hat{r}, a | I, O, S; \theta_t) \quad (4)$$

This procedure encourages the model to internalize structured reasoning patterns and align its final judgments with both synthetic rationales and human preferences.

## 5. Experiments

### 5.1. Benchmarks and Metrics

We evaluate AUTORATER and baselines, including the base VLMs and a specialist, on the following three evaluation scenarios. We also summarize the corresponding benchmarks information in Table 1.

**Image editing (pairwise comparison):** We use the test set of PhotoShopRequest (PSR) dataset [25], which contains real human queries and human edits from Reddit. In this dataset, an input-output set  $(I, O)$  can have multiple human annotations. Each annotation is one of three options: two options of which among the two edited images in  $O$  is better, and one option of ‘tie’. Thus, we aggregate all human annotations for each of  $(I, O)$ . We only keep an

input-output set  $(I, O)$  if the mode<sup>2</sup> vote accounts for at least 50% of all annotations for the set and has at least 2 more votes than the second mode vote. We use the mode vote as the final annotation for the input-output set. Also, as the number of samples with the final annotation ‘tie’ is 8 out of 452, we do not consider this case in this work. Overall, after the aggregation and filtering, there are 444 tuples  $(I, O, a)$ .

**Image generation (pairwise comparison):** We employed the test set of ImageReward [30]. For a generation request, there are 4 to 9 generated images, and they are sorted into 5 ranking bins from 1 to 5. We follow the conventional setting in this line of work, where we consider all different-rank image pairs from each sample. This results in 6,399 pairs in total.

**Image generation (point-wise rating):** The test set of ImageReward [30] is used in a different setting. For each of the generated images, we use the overall score as the ground-truth rating, yielding 2,720 instances in total. We also consider the test set of AIGIQA-30K [15] with 4,000 samples. Each sample consists of a generation request and a generated image, and a human-annotated score. We treat the human-annotated score as the ground-truth rating.

### 5.2. Experiment Results for ICL Methods

**Model Setup:** For AUTORATER, we consider two VLMs as the base models: GPT-4o and Gemini-2.5-Pro. Both models are strong VLMs that achieve high performance on multiple vision-language benchmarks. We set the temperature to 0 for both models to ensure deterministic outputs. Also, we use *Likert-10 scale* in case of pointwise preference. For image assessment specialists, we use the official open-source implementations with default configurations. We show VLM judge prompts for each evaluation scenario and the way we organize specialists’ output in the ???. We compare AUTORATER with strong baselines: the base VLMs (GPT-4o and Gemini-2.5-Pro) themselves, and three image preference models ImageReward [30], Q-SiT [35], and the recently released HPSv3 [17]. All empirical results are shown in Table 2.

**Results:** GPT-4o is relatively strong in pointwise rating with high Pearson correlations 0.7233 on AIGIQA-30K, while Gemini-2.5-Pro is good in pairwise comparison, especially in PSR with 78.15% accuracy. Meanwhile, the preference model HPSv3 yields comparable results with GPT-4o and Gemini-2.5-Pro, with highest accuracy of 66.8% on ImageReward (Pair) and highest Pearson correlation 0.782 on AIGIQA-30K among the baselines. On the other hand, ImageReward and Q-SiT falls short in two new benchmarks, PSR and ImageReward (pointwise). Their accuracy on PSR is approximately or less than random guessing, leaving a large gap behind a general VLM

<sup>2</sup>The vote that appears most frequently.

Model	PSR $\uparrow$	ImageReward (Pair) $\uparrow$	ImageReward (Point) $\uparrow$	AIGIQA-30K $\uparrow$
HPSv3	52.03	<b>66.80</b>	0.4524	0.7820
ImageReward	50.90	65.10	0.3575	0.6938
QSiT	22.35	55.13	0.3441	0.7526
GPT-4o	64.41	57.66	0.3176	0.7233
+ Auto	63.51	58.60	0.3861	0.7932
+ All	66.67	57.98	0.3939	0.8047
+ Grid-Search	67.79 <sub>DF,GD,QS</sub>	58.81 <sub>AE,GD,PA</sub>	0.4084 <sub>AE,PA,QS</sub>	0.8048 <sub>AE,QS</sub>
Gemini-2.5-Pro	78.15	61.84	0.3150	0.6430
+ Auto	76.58	61.02	0.3392	0.6429
+ All	74.55	60.82	0.3446	0.6578
+ Grid-Search	<b>78.38</b> <sub>DF,QS</sub>	62.64 <sub>AE,PA,QS</sub>	0.3612 <sub>QS</sub>	0.6766 <sub>QS</sub>
InternVL3-8B	49.55	59.22	0.4186	0.7240
+ CoT finetuning w/o Tools	54.27	59.51	0.3855	0.7669
+ CoT finetuning w/ all Tools	58.56	60.53	<b>0.4673</b>	<b>0.8533</b>

Table 2. Experiment results on four benchmarks. For pairwise comparison, we report accuracy (%). For pointwise rating, we report Pearson correlation. The best tool combinations for Grid-Search are subscripted behind the reported numbers. **Bold-faced** and *italic* numbers respectively indicate the best and second best performance in each benchmark.

Gemini-2.5-Pro, while its Pearson correlation of 0.35% of both models on ImageReward (pointwise) still has room for improvement, as what a variant of AUTORATER can bring.

**Potential of AUTORATER and improvement gaps in different scenarios:** For base model GPT-4o, pointwise correlations improve substantially when augmenting with specialists: on ImageReward (pointwise), the best combination (with AesExpert, PAL4VST, Q-SiT) reaches 0.4084, compared to 0.3176 for GPT-4o alone and 0.3939 for GPT-4o with All specialists. On AIGIQA-30K, using All specialists yields 0.8047 vs. 0.7233 for the base model. The improvement brought by specialists is less on ImageReward (pairwise) and PSR, up to 1% and 3% respectively. A similar trend can also be observed for the base model Gemini-2.5-Pro. While there is no large improvement or even a slight decrease in ImageReward (pairwise) and PSR, AUTORATER achieves the highest Pearson correlation on ImageReward (pointwise) at 0.3612 and on AIGIQA-30K at 0.6766 with the augmentation of the specialist Q-SiT.

In pointwise rating scenarios, we observe that Q-SiT often appears in the best combinations for both base VLMs. Given a base VLM, AUTORATER with individual Q-SiT has at least 3% gap to any other variants with one specialist. It is reasonable as Q-SiT is trained for image quality scoring and achieves SoTA performance on multiple image assessment benchmarks. Q-SiT provides calibrated, low-level quality cues that directly inform scalar judgments. Interestingly, each baseline among GPT-4o and Gemini-2.5-Pro together with Q-SiT forms a baseline that is stronger than these baselines alone. Thus, *we posit that having a strong specialist functionally related to the task would benefit base VLMs to go beyond both of these individual baselines.*

In pairwise rating scenarios, to our best knowledge, there is a limited number of specialized image assessment tools that compare two images. In our experiments, all are tools considered for pairwise comparison scenarios do not directly function as a pairwise evaluator. The insights from these tools may not have explicit effect on the final judgment. Our later qualitative analyses elaborate on the situation. Thus, *we posit that in order to better augment VLMs in pairwise comparison, specialists who directly compare images are desired.*

Further analyses on tool selection mechanisms will be provided in the ??.

### 5.3. Experiment Results for Finetuning Methods

**Model Setup:** We consider open-source model family InternVL [27] as 1) the model for data generation and 2) the pretrained model for finetuning because of its superior performance in public benchmarks. In particular, we use the InternVL3.5-14B to generate high-quality reasoning data following the method in Section 4.3. We then train the InternVL3-8B[36] model with the synthetic data. We generate specialist-guided rationale based on labels from the training set of ImageReward [30], AIGIQA-30K [15], and EditReward [28], resulting in 42k samples for image generation evaluation and 13k samples for image editing evaluation in total. The hyperparameter details are provided in ??.

**Results:** With Chain-of-Thought (CoT) finetuning, InternVL improves across all metrics—especially compared to the base non-finetuned model. Yet, the largest leap appears when CoT finetuning is aided with specialists’ assessments, reaching 58.56% accuracy on PSR, 60.53% accuracy

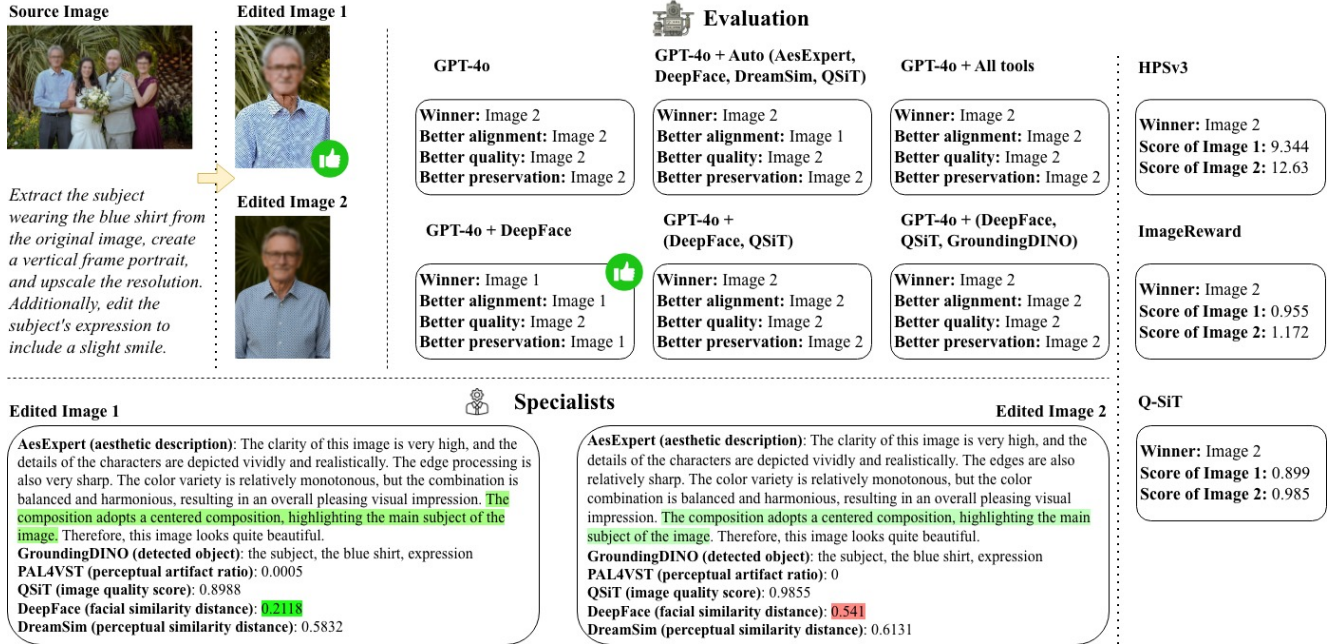


Figure 3. A sample from PSR benchmark. GPT-4o fails to identify the preservation of facial identity in both edited images, thus, makes a wrong judgment. With the help of DeepFace, AUTORATER can identify the preservation in Edited Image 1 and correct its judgment. However, when more tools are added, the extra information makes the assessment noisy, and AUTORATER fails again.

on ImageReward (Pair), 0.4673 Pearson correlation on ImageReward (Point), and an impressive 0.8533 Pearson correlation on AIGIQA-30K. This final configuration outperforms all other models in both ImageReward (Point) and AIGIQA-30K, showing that the synergy between CoT reasoning and tool integration can yield substantial improvements in perceptual quality evaluation. Nonetheless, finetuned InternVL is still lagging behind GPT-4o and Gemini-2.5-Pro in PSR. We attribute this to the lack of photo-realistic image editing data in our training data<sup>3</sup>, thus InternVL has not fully learned the task.

Comparing our finetuned InternVL with HPSv3 [17], a State-of-The-Art image preference model with a similar size, it outperforms HPSv3 on the pointwise preference for image generation. In AIGIQA-30K, the gap is 7%. Meanwhile, in PSR and ImageReward (Pair), each model is better in a benchmark, with a similar performance on average. Undoubtedly, our InternVL achieves a better overall performance than HPSv3, while only using 55k samples comparing to millions samples of HPSv3.

**Ablation Study** Table 3 shows the ablation study with finetuning and inferencing the open-source model InternVL3-8B. When the model is fine-tuned with tools and also uses them during inference (Full), it achieves the highest score, while the base model (5) yields the lowest Pearson. Fine-

<sup>3</sup>Reddit did not grant permission to use Reddit data for AI training, and the PSR dataset did not include large-scale human preference besides a few hundred samples for evaluation

Setup	FT	FT w/ tools	Infer w/ tools	AIGIQA-30K ↑
<b>Full</b>	✓	✓	✓	<b>0.8533</b>
1	✓	✓		0.7981
2	✓		✓	0.7719
3	✓			0.7669
4			✓	0.7740
5				0.7240

Table 3. Ablation study on AIGIQA-30K using InternVL3-8B variants with the inclusion of finetuning, finetuning with tools, and inference with tools.

tuning with tools but disabling them at inference (1) still improves performance over standard fine-tuning, indicating that tool-aware training imparts more robust reasoning even without tool access at test time. Introducing tools only at inference without tool-aware fine-tuning (2) results in more modest and inconsistent gains, suggesting that while tools provide auxiliary benefits, a model not trained to use them cannot fully capitalize on their capabilities. Standard finetuning without any tools (3) remains the weakest among the fine-tuned configurations, demonstrating the limited impact of parameter-only learning in this task compared to tool-augmented approaches. Having access to the tools at inference time without any finetuning (4) offers a gain over pure zeroshot ICL (5) however, significantly lower than finetuning with tool awareness (Full). Overall, the study shows

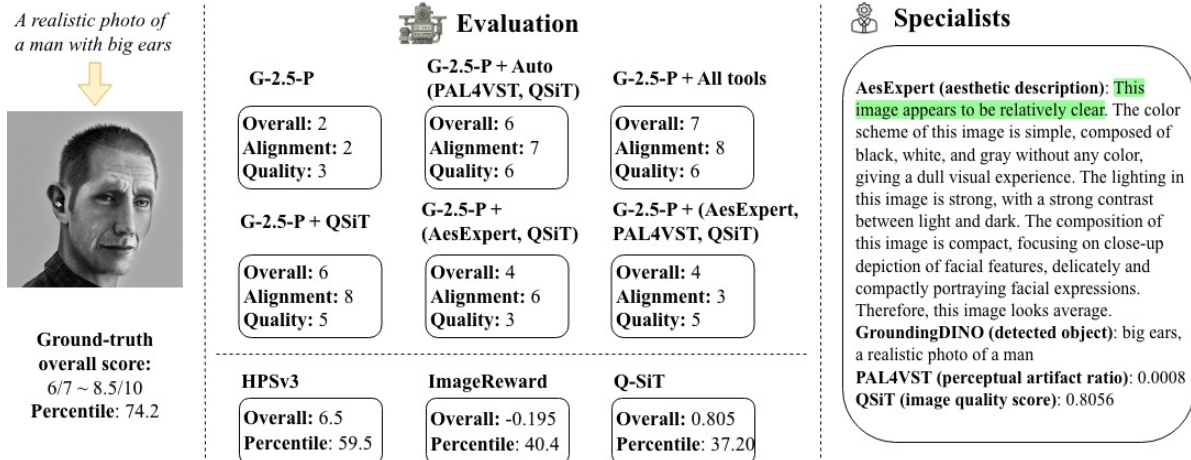


Figure 4. Qualitative analysis on ImageReward (pairwise) benchmark and Gemini-2.5-Pro base model (denoted as **G-2.5-P** in the figure). We use *Likert-10 scoring scale* for this pointwise preference scenario. Gemini-2.5-Pro alone underestimates the quality of the generated image, giving a low score of 2. With the help of Q-SiT, every AUTORATER variant gives a more ground-truth-aligned score (4 or 5). However, Gemini-2.5-Pro remains deficient in absolute scoring, sometimes assigning lower scores even with specialists’ assessments, which indicates decent alignment and quality of the image.

that while tools generally improve performance, the greatest benefit arises when the model is trained to use tools for reasoning and is allowed to access the tool during inference.

#### 5.4. Qualitative Analysis

To illustrate the potential in augmenting VLMs with specialist tools for pairwise comparison, we analyze a representative sample from the PSR benchmark in Figure 3. In this case, GPT-4o alone struggles to accurately assess the facial/identity preservation, which is pivotal for image editing tasks involving human subjects. When DeepFace is introduced as a specialist, the assessment improves: GPT-4o references facial similarity scores and acknowledges a closer match to the original subject in one edited image. Together with text-image alignment, the edited image is correctly judged by GPT-4o + DeepFace as the better edit. However, when augmenting GPT-4o with more specialists, the judgment converges back to an incorrect decision as baseline GPT-4o. It suggests that adding further tools sometimes introduces noise, diluting the clarity of the assessment and causing GPT-4o to misjudge preservation. This highlights the need for specialists which explicitly compare images and provide the decision’s rationale to better guide VLM judges.

In the pointwise rating scenario, as shown in Figure 4, integrating a strong specialist like Q-SiT with Gemini-2.5-Pro consistently improves alignment with human ratings. All AUTORATER variants give scores with higher alignment with human preference than that of the base Gemini-2.5-Pro, demonstrating the benefit of functionally relevant specialists for scalar judgments. Gemini-2.5-Pro remains deficient in absolute scoring, sometimes assigning lower

scores even when a specialist’s assessment indicates decent alignment and quality of the image. Subscores from Q-SiT provide valuable low-level cues, but the VLM’s final rating may still underrepresent image quality. These observations suggest that while specialist augmentation enhances VLM performance, further refinement in score calibration and specialist-VLM interaction is needed for optimal pointwise assessment.

#### 6. Conclusion

In this paper, we explore the combination of VLMs and specialized image assessment tools to improve automatic preference assessment for image generation and editing models. We design a system AUTORATER, consider three specialist selection mechanisms to provide extra context to VLM judges, and investigate both in-context learning and finetuning approaches. Through extensive experiments on multiple benchmarks, we demonstrate that integrating specialist tools can enhance preference alignment over strong VLM baselines. In addition, our finetuning approach demonstrates that the VLM judge can learn to reason with specialist assessments, further boosting alignment with human preferences. We conclude the paper with qualitative analyses illustrating cases where specialist assessments effectively aid zero-shot VLM judges. We also analyze failure cases where they introduce noise or fail to capture key details. We hope that our work sheds light on the potential of combining VLMs with specialist models for reliable image preference assessment.

## References

- [1] Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Leria HUANG, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. MJ-bench: Is your multimodal reward model really a good judge? In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. 3
- [2] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 1
- [3] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977. 5
- [4] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, pages 50742–50768, 2023. 4
- [5] Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, et al. Mllm-bench: evaluating multimodal llms with per-sample criteria. *arXiv preprint arXiv:2311.13951*, 2023. 1, 3
- [6] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 1, 3
- [7] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 3
- [8] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhen-guo Li, and Xihui Liu. T2i-compench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:3563–3579, 2023. 3
- [9] Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *ACM Multimedia 2024*, 2024. 3
- [10] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024. 3
- [11] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4324–4333, 2024. 2
- [12] Yoonjeon Kim, Soohyun Ryu, Yeonsung Jung, Hyunkoo Lee, Joowon Kim, June Yong Yang, Jaeryong Hwang, and Eunho Yang. Preserve or modify? context-aware evaluation for balancing preservation and modification in text-guided image editing. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23474–23483, 2024. 3
- [13] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: an open dataset of user preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 1, 2, 3
- [14] Max W.F. Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhui Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *ArXiv*, abs/2312.14867, 2023. 3
- [15] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Aigiqa-20k: A large database for ai-generated image quality assessment. In *CVPR Workshops*, pages 6327–6336, 2024. 2, 5, 6
- [16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2023. 3
- [17] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025. 2, 3, 5, 7
- [18] OpenAI. Gpt-4 technical report. 2023. 3
- [19] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, J. Heikkila, and Shin’ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14277–14286, 2023. 3
- [20] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 1
- [21] Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. Ice-bench: A unified and comprehensive benchmark for image creating and editing. *arXiv preprint arXiv:2503.14482*, 2025. 3
- [22] Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. The photographer’s eye: Teaching multimodal large language models to see, and critique like photographers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24807–24816, 2025. 2
- [23] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations*

- in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020. 4
- [24] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8871–8879, 2023. 2
- [25] Mohammad Reza Taesiri, Brandon Collins, Logan Bolton, Viet Dac Lai, Franck Dernoncourt, Trung Bui, and Anh Totti Nguyen. Understanding generative ai capabilities in everyday image editing tasks, 2025. 2, 5
- [26] Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Understanding and evaluating human preferences for ai generated images with instruction tuning, 2024. 3
- [27] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6
- [28] Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025. 1, 6
- [29] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2, 3
- [30] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935, 2023. 1, 2, 3, 5, 6
- [31] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023. 2
- [32] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7579–7590, 2023. 4
- [33] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024. 2, 3
- [34] Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-bench+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:10404–10418, 2024. 2
- [35] Zicheng Zhang, Haoning Wu, Ziheng Jia, Weisi Lin, and Guangtao Zhai. Teaching llms for image quality scoring and interpreting, 2025. 1, 3, 4, 5
- [36] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yanan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 6