CAPTURE WITH PHOTOENCODER: PHOTOGRAPHIC UNDERSTANDING WITH REPRESENTATION ENHANCED VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

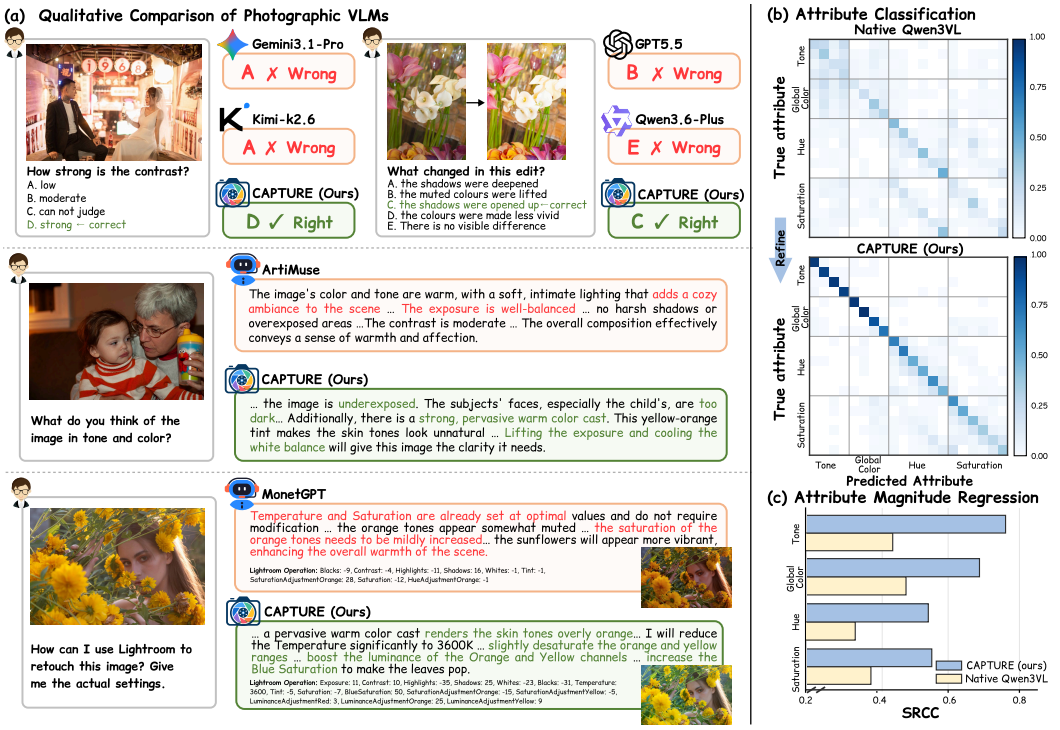


Figure 1: (a) Qualitative comparisons of existing VLMs and CAPTURE across photographic perception, aesthetic understanding, and retouching assistance. (b) Confusion matrices from the PDP for attribute classification, comparing Native Qwen3VL with CAPTURE. (c) Spearman rank correlation coefficients (SRCC) for per-attribute change-magnitude regression under the PDP. Higher values indicate stronger rank agreement between predicted and measured change magnitudes.

ABSTRACT

Vision-language models (VLMs) could be powerful tools for understanding and retouching photographs, yet their ability to perceive fine-grained photographic attributes, such as brightness and color temperature, remains limited. Using the **Photographic Diagnostic Probe**, we reveal that existing VLM vision encoders inadequately capture these photographic attributes. To address this limitation, we introduce **PhotoEncoder**, which structures visual representations around photographic attributes and their controlled transitions while preserving pretrained semantic representations. We further construct a dataset of 5M vision-language pairs and 81K instruction-tuning examples derived from editing records to support training. By integrating PhotoEncoder into the language model and applying instruction tuning, we build **CAPTURE** for photographic perception, aesthetic understanding, and retouching assistance. Experiments show that CAPTURE better distinguishes subtle tonal and color changes, enabling more reliable aesthetic judgments and editing decisions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Developing computational assistants for photography has attracted increasing interest in both research and real-world applications (Qi et al., 2025; Lin et al., 2025; OpenAI, 2026; Google, 2026; Zeng et al., 2026; Zhu et al., 2026). Recent advances in vision-language models (VLMs) have opened new possibilities for photographic assistance through natural-language interaction, enabling systems that critique image aesthetics and analyze visual qualities (Huang et al., 2024a; Qi et al., 2025; Du et al., 2026; Cao et al., 2026), and translate user intent into executable retouching operations (Chen et al., 2025a; Dutt et al., 2025; Lin et al., 2025; 2026). Despite these advances, existing VLMs still exhibit unreliable photographic understanding. As shown in Fig. 1(a), they mistake strong contrast for low contrast, fail to recognize lifted shadows, and describe an underexposed image with a warm color cast as well-exposed and pleasantly lit. Recent efforts have produced aesthetic instruction-tuning datasets (Huang et al., 2024a; Qi et al., 2025; Zhou et al., 2024; Cao et al., 2026), providing VLMs with supervision from aesthetic critiques and expert analyses. However, even after photographic post-training, these models still struggle to distinguish fine-grained tonal and color attributes. These remind us of a possibility: do vision encoders limit current VLMs' ability to perceive fine-grained photographic attributes?

To investigate the visual representations, we adapt a diagnostic technique called **Photographic Diagnostic Probe (PDP)**, which isolates visual representations from language reasoning and directly evaluates whether photographic attributes are encoded in VLM representations. In Fig. 1(b), the confusion matrix for the native Qwen3VL representation places substantial mass away from the diagonal, indicating frequent confusion among these photographic attributes. In Fig. 1(c), the per-attribute SRCCs are generally low and uneven, indicating weak correspondence between the predicted and measured change strengths. Together, these observations indicate that native vision representations inadequately capture fine-grained photographic attributes and the magnitudes of their changes.

Therefore, we introduce **PhotoEncoder**, a vision encoder explicitly adapted to represent fine-grained photographic attributes. PhotoEncoder is trained to capture both the tonal and color state of each image and how these attributes change after editing, while preserving the semantic content of the pretrained encoder. Building on PhotoEncoder, we construct **CAPTURE** by integrating the enhanced vision encoder back into the language model and performing supervised fine-tuning on diverse photographic-assistant tasks. This enables CAPTURE to interpret fine-grained attributes from images, understand user editing intents, and translate them into 41 actionable Lightroom operations, such as increasing exposure, reducing highlights, and warming the color temperature.

To support photographic assistance, we construct a photographic dataset containing 5M vision-language pairs for PhotoEncoder adaptation and 81K photographic instruction-tuning samples for VLM training. We further establish a 3K-example benchmark covering photographic perception, aesthetic understanding, and retouching assistance.

2 RELATED WORK

Aesthetic and Low-Level Visual Instruction Tuning. Recent works extend VLM instruction tuning to low-level visual perception, image quality, and aesthetic understanding (Wu et al., 2024; You et al., 2024; 2025; Li et al., 2025; Huang et al., 2024a; Cao et al., 2026; 2025). Photographic editing systems further train multimodal models to analyze images, recommend edits, or invoke executable editing operations (Yao et al., 2026; Wu et al., 2026; Lin et al., 2025; 2026; Zhang & Yang, 2026). These approaches primarily improve photographic capabilities through instruction tuning, task-specific supervision, or encoder fusion, without explicitly adapting the native vision encoder for fine-grained tonal and color variations. Meanwhile, existing datasets (You et al., 2024; Huang et al., 2024b; Zhou et al., 2024; Qi et al., 2025; Lin et al., 2025) typically focus on low-level visual understanding, aesthetic assessment, or editing operations in isolation, leaving photographic supervision fragmented across different task families. We directly adapt the native vision encoder and construct unified photographic supervision from controlled edits and expert retouching records for perception, critique, and retouching.

Visual Representations for Vision and Aesthetic Understanding. Most VLMs build on general-purpose pretrained visual encoders such as CLIP and SigLIP (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025), which are primarily trained for broad visual-semantic alignment rather than

fine-grained photographic understanding (Qi et al., 2025). Several works explore specialized visual representations for image quality and aesthetics (Wang et al., 2023; He et al., 2023b; Behrad et al., 2025; Jia et al., 2026). More specialized studies target color-related representations, including image color aesthetics, color-temperature assessment, and cross-image color-difference interpretation (He et al., 2023a; Cheng et al., 2026; Wang et al., 2025b). These approaches show that visual representations can be specialized for perceptual and aesthetic objectives, but generally focus on particular tasks or visual properties. We adapt the native VLM encoder to better represent multiple tonal and color properties and their editing-induced changes.

Diagnosing Fine-Grained Visual Representations. A growing body of work evaluates fine-grained visual and color perception in VLMs (Tong et al., 2024; Fu et al., 2024; Liang et al., 2025; Chiu et al., 2026). Beyond behavioral evaluation, diagnostic probing provides a direct way to examine what information is encoded in frozen representations (Alain & Bengio, 2016; Hewitt & Liang, 2019). We use the PDP to directly probe tonal and color representations in frozen vision encoders.

3 Probing Tonal and Color Representations

We adapt the PDP to assess whether vision encoder representations capture fine-grained photographic attributes and the magnitude of their changes. To this end, the PDP is trained on two complementary tasks: *attribute classification*, which identifies the photographic attribute that changes within an image pair, and *magnitude regression*, which estimates the strength of that change.

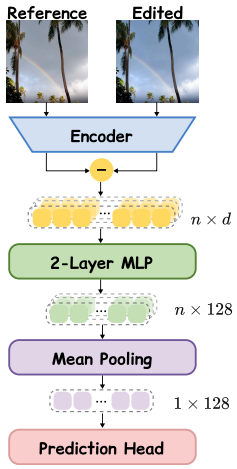


Figure 2: PDP architecture.

Photographic Diagnostic Probe. We define the probe labels using 26 photographic attributes measured from the rendered images: four tonal attributes (brightness, contrast, highlight level, and shadow level), four global color attributes (temperature, tint, global saturation, and chroma distribution), and hue, saturation, and lightness within each of six hue ranges. Their measurement definitions are provided in Appendix A.1. We construct controlled reference–edited pairs, as illustrated in Fig. 2; the frozen encoder produces visual tokens for each image, and their position-aligned difference is used as the probe input. We apply a two-layer MLP followed by mean pooling and a linear head, training one 26-way classifier for attribute identification and 26 attribute-specific regressors for magnitude estimation.

Empirical Findings. As summarized in Tab. 4, we report classification accuracy for attribute classification and mean Spearman rank correlation coefficient (SRCC) across attributes for magnitude regression. The results reveal limitations in both attribute identification and change-magnitude estimation for the native Qwen3VL encoder (Bai et al., 2025a). It achieves 26.69% classification accuracy across the 26 attributes and a mean SRCC of 0.376 for magnitude regression. In Fig. 1(b), the native encoder frequently confuses distinct attributes; for example, 25.15% of brightness changes are classified as contrast changes. For magnitude estimation, Fig. 1(c) reports averaged SRCCs of 0.441 for tonal attributes and 0.336 for hue attributes, indicating limited rank consistency between predicted and measured change magnitudes. Our PhotoEncoder adaptation improves both diagnostics, yielding clearer attribute separation and stronger magnitude rank consistency in Fig. 1(b,c).

4 Method

Based on the observations in Section 3, we adapt the native vision encoder into PhotoEncoder by supervising both the photographic attributes of individual images and their changes across edits, while preserving the encoder's pretrained semantic representations. Building on PhotoEncoder, we further develop CAPTURE through photographic instruction tuning, enabling photographic perception, aesthetic understanding, and retouching assistance, as illustrated in Fig. 3.

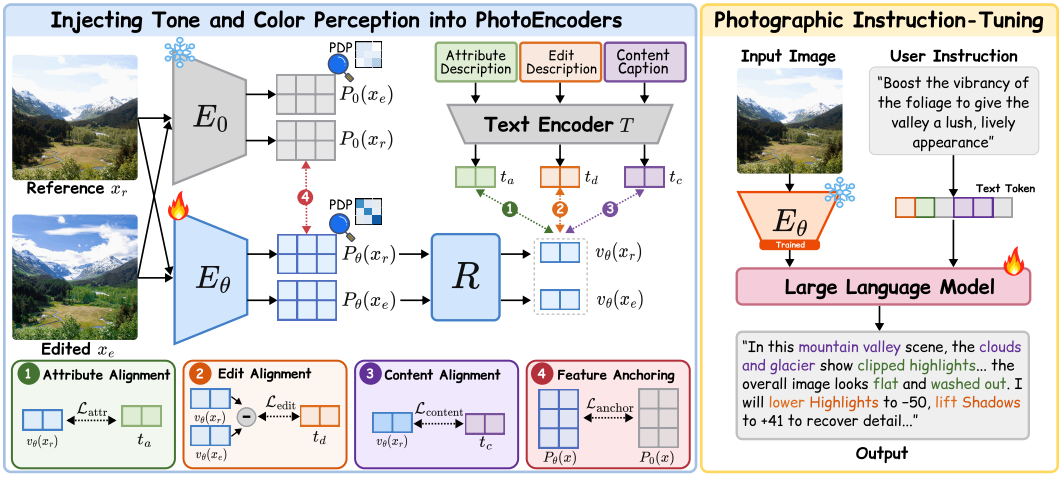


Figure 3: Overview of PhotoEncoder adaptation and photographic instruction tuning. PhotoEncoder is adapted with attribute alignment, edit alignment, feature anchoring, and content alignment. The adapted PhotoEncoder is then integrated into the LLM and instruction-tuned for photography.

4.1 Injecting Tone and Color Perception into PhotoEncoder

The diagnostic findings in Section 3 motivate adaptation at the representation level, where fine-grained tonal and color changes remain difficult to distinguish and quantify. Our goal is to organize the native encoder's representation space around photographic attributes and their controlled changes, while minimizing disruption to its pretrained representations (Lee et al., 2024).

Inputs and Representations. Each training sample contains a reference image x_r and an edited image x_e of the same scene, where x_e is obtained by varying a controlled photographic attribute. For each image, we have measured tonal and color attribute levels; for each pair, we additionally know the changed attribute and its change direction. A content caption shared by the pair describes scene semantics independent of the edit. Let E_θ denote the adapted vision encoder, E_0 its frozen initialization, R a backbone-specific image-level readout, and T a frozen text encoder. $P_\theta(x)$ and $P_0(x)$ respectively denote the patch representations produced by E_θ and E_0. We define

$$v_\theta(x) = \text{normalize}(R(P_\theta(x))), \; \Delta v_\theta = \text{normalize}(v_\theta(x_e) - v_\theta(x_r)), \; t_p = \text{normalize}(T(p)), \tag{1}$$

where $x \in \{x_r, x_e\}$ and p is a supervision phrase. We use t_a, t_d, and t_c for attribute descriptions, edit descriptions, and content captions, respectively.

Depending on the backbone, R either aggregates patch tokens or projects a global CLS token into the text embedding space. For alignment objectives, we use the pairwise sigmoid formulation of SigLIP (Zhai et al., 2023). For objective j, the similarity logit is $S_j(z,t) = s_j z^\top t + b_j$, where s_j and b_j are learned scale and bias.

Overall Objective. We optimize four complementary objectives:

$$\mathcal{L}_{\text{core}} = \lambda_{\text{attr}}\mathcal{L}_{\text{attr}} + \lambda_{\text{edit}}\mathcal{L}_{\text{edit}} + \lambda_{\text{content}}\mathcal{L}_{\text{content}} + \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}}. \tag{2}$$

Attribute alignment $\mathcal{L}_{\text{attr}}$ supervises which attribute levels an image representation encodes, while edit alignment $\mathcal{L}_{\text{edit}}$ encourages representation differences to distinguish the changed attribute and follow its measured change direction consistently across images. Content alignment $\mathcal{L}_{\text{content}}$ maintains correspondence between image representations and scene descriptions, while feature anchoring $\mathcal{L}_{\text{anchor}}$ limits drift from the initial patch representations. Together, these objectives balance sensitivity to photographic attributes with preservation of the pretrained representation space.

Attribute Alignment. A useful photographic representation should encode not only scene content, but also the absolute tonal and color state of an image. We therefore organize image representations according to measured photographic attribute levels, so that images with different brightness, contrast, saturation, or color states become distinguishable in the embedding space. For each attribute, we measure the image and assign it to a discrete level represented by a textual description. Let a

index the attribute-level descriptions and t_a their text embeddings. We align $v_\theta(x)$ with t_a:

$$\mathcal{L}_{\text{attr}} = -\sum_{x\in\{x_r,x_e\},a} \log\sigma(q_{x,a}S_{\text{attr}}(v_\theta(x),t_a))\,, \tag{3}$$

where $q_{x,a} = 1$ when a describes the measured level of x, $q_{x,a} = 2\alpha - 1$ for an adjacent level, and $q_{x,a} = -1$ otherwise, with $0 < \alpha < 0.5$. Treating neighboring levels as weaker negatives reflects the ordinal structure of photographic attributes, where adjacent levels are less distinct than distant ones. This objective encourages the visual space to preserve attribute identity and level relations.

Edit Alignment. Photographic assistance requires not only recognizing the tonal and color state of a single image, but also understanding how it changes after an edit. Such change-aware representations are important for comparing editing outcomes, assessing whether an adjustment produces the intended effect, and supporting retouching decisions. We therefore complement image-level attribute supervision with transition-level supervision, encouraging controlled edits to induce structured and consistent movements in the representation space. Specifically, we align Δv_θ with an edit description t_d specifying both the changed attribute and its measured direction (Sam et al., 2024):

$$\mathcal{L}_{\text{edit}} = -\sum_{d} \log\sigma(y_d S_{\text{edit}}(\Delta v_\theta, t_d))\,, \tag{4}$$

where $y_d = 1$ when d matches both the changed attribute and its direction, and $y_d = -1$ otherwise. This objective encourages edits that induce the same photographic change to follow consistent directions in the representation space across different images.

Content Alignment. Directly restructuring the visual space around photographic attributes may disturb semantic information learned during pretraining. We therefore preserve the pretrained representation at both the image and patch levels. Let t_c denote the text embeddings. We define

$$\mathcal{L}_{\text{content}} = -\sum_{x\in\{x_r,x_e\}} \sum_{c} \log\sigma(y_{x,c}S_{\text{content}}(v_\theta(x),t_c))\,, \tag{5}$$

where $y_{x,c} = 1$ when c is the content caption associated with x, and -1 for mismatched captions. This encourages images of the same scene to retain their semantic correspondence.

Feature Anchoring. At the patch level, $\mathcal{L}_{\text{anchor}}$ minimizes the cosine distance between corresponding patch features in $P_\theta(x)$ and $P_0(x)$:

$$\mathcal{L}_{\text{anchor}} = \mathbb{E}_{x\in\{x_r,x_e\}} \frac{1}{N_x}\sum_{i=1}^{N_x} [1 - \cos(p_{\theta,i}(x), p_{0,i}(x))]\,, \tag{6}$$

where $p_{\theta,i}(x)$ and $p_{0,i}(x)$ are the corresponding i-th patch features and N_x is the number of patches. Feature anchoring prevents visual representations from drifting excessively during adaptation.

4.2 Photographic Instruction-Tuning

To translate the enhanced visual representations into practical photographic assistance, we perform supervised fine-tuning using the adapted PhotoEncoder. We organize the instruction data into three task sets: **photographic perception**, **aesthetic understanding**, and **retouching assistance**. Each example contains one or more related images, a user instruction, and a target response. The three task families provide complementary supervision. Photographic perception covers photographic attributes and their changes, aesthetic understanding covers aesthetic quality and style assessment, and retouching assistance covers editing recommendations and Lightroom operations. We jointly train on these task sets with standard autoregressive supervised fine-tuning. During fine-tuning, the adapted PhotoEncoder remains frozen, while the visual merger and language model are optimized to generate task-specific responses conditioned on the images and user instructions. This transfers the fine-grained photographic representations to downstream capabilities in photographic perception, aesthetic understanding, and retouching assistance.

5 EXPERIMENTS

5.1 DATASET AND BENCHMARK

To support the training of CAPTURE, we construct a photographic dataset from controlled edits and expert editing records in MIT-Adobe FiveK (Bychkovsky et al., 2011) and PPR10K (Liang et al., 2021). It contains approximately 5M vision–language pairs for PhotoEncoder adaptation, 81K photographic instruction examples, and a 3K-example benchmark.

PhotoEncoder Adaptation. We generate controlled reference–edited pairs over 26 photographic attributes. The supervision covers three forms: image–attribute-level pairs for attribute alignment, representation-change–direction pairs for edit alignment, and image–caption pairs for preserving pretrained representations and reducing semantic drift during adaptation.

Photographic Instruction Data. The instruction data span three task families. 1) Photographic perception includes *attribute assessment* and *transition assessment*, both derived from the controlled pairs used for PhotoEncoder adaptation. 2) Aesthetic understanding includes *aesthetic critique* and *style comparison*. We construct these tasks through knowledge inversion (Li et al., 2026) with Gemini3.1-Pro (Google, 2026), using original images, expert-retouched results, and editing parameters to recover the underlying aesthetic knowledge. 3) Retouching assistance includes *editing recommendation* and *Lightroom parameter prediction*. Editing recommendations are likewise obtained through knowledge inversion, while parameter prediction directly uses recorded expert editing parameters as supervision. Details are provided in Appendix A.5.

Comparison with Existing Datasets. Tab. 1 highlights the broader supervision coverage of our dataset. Unlike prior datasets (Wu et al., 2024; Qi et al., 2025; Cao et al., 2026; Lin et al., 2025) that primarily target individual downstream capabilities, our construction jointly supports representation-level PhotoEncoder adaptation and instruction tuning across photographic perception, aesthetic understanding, editing recommendation, and Lightroom parameter prediction. In particular, the controlled image pairs provide explicit supervision over photographic attributes and their changes, while the instruction data extend these representations to complementary photographic tasks. This unified coverage connects fine-grained visual representation learning with downstream photographic understanding and editing within a single data framework.

Table 1: Comparison of task coverage across datasets. Datasets with * are currently not available.

Dataset	Enc. Adapt.	Photo. Percep.	Aesthetic Underst.	Edit Recom.	LR Param.
Q-Instruct	✗	✓	✓	✓	✗
AesMMIT	✗	✗	✓	✓	✗
ArtiMuse	✗	✓	✓	✗	✗
PhotoCritique*	✗	✓	✓	✓	✗
MMArt	✗	✗	✗	✓	✓
Ours	✓	✓	✓	✓	✓

Metrics. For the PDP, we report classification accuracy for attribute identification and Spearman rank correlation coefficient (SRCC) across attributes for magnitude regression. SRCC measures the rank consistency between predicted and measured change magnitudes, indicating whether the representation preserves the relative strength of changes within each attribute. For benchmark evaluation, photographic perception is evaluated with multiple-choice questions, for which we report accuracy. For open-ended aesthetic understanding and retouching instruction generation, we adopt LLM-as-a-judge evaluation (Zheng et al., 2023), using Qwen3.6-27B (Qwen Team, 2026b) with task-specific reference answers and scoring rubrics. For Lightroom parameter prediction, we render predicted edits and compare them against expert edits with PSNR and SSIM (Wang et al., 2004).

5.2 EXPERIMENTAL SETUP

Encoder Adaptation. We adapt the Qwen3VL-8B-Instruct vision encoder (Bai et al., 2025a) and InternViT (Wang et al., 2025a) using LoRA (Hu et al., 2022) ($r = 16$, $\alpha = 32$, dropout 0.05). Image–text supervision uses a SigLIP2 MAP readout for Qwen3VL and a CLIP CLS readout for InternViT, each paired with its corresponding frozen text encoder. For Qwen3VL, we train for 8 epochs on 8 NVIDIA A6000 GPUs using AdamW with a learning rate of 10^{-4} and weight decay of 0.01. The core loss weights are $\lambda_{\text{attr}} : \lambda_{\text{edit}} : \lambda_{\text{anchor}} : \lambda_{\text{content}} = 1 : 1 : 0.1 : 0.1$.

Photographic Instruction-Tuning. Following vision-encoder adaptation, we replace the original Qwen3VL-8B-Instruct vision encoder with our PhotoEncoder and perform supervised fine-tuning on the photographic instruction data. During instruction tuning, the PhotoEncoder is kept frozen,

Table 2: Performance of VLMs on photographic perception and aesthetic understanding, evaluated by accuracy (%) and task-specific LLM-judge scores (mean ± standard deviation), respectively.

Model	Photographic Perception (Acc. % ↑)		Aesthetic Understanding (Judge ↑)	
	Attribute Assessment	Transition Assessment	Aesthetic Critique	Style Comparison
Proprietary Models				
GPT5.5 (OpenAI, 2026)	48.61	72.39	5.327 ± 0.007	**7.673 ± 0.006**
Gemini3.1-Pro (Google, 2026)	45.68	73.47	1.421 ± 0.006	5.658 ± 0.014
Qwen3.6-Plus (Qwen Team, 2026c)	46.56	41.48	3.345 ± 0.010	4.221 ± 0.007
Open Source Models				
LLaVA-OneVision2-8B (An et al., 2026)	32.21	8.02	0.874 ± 0.004	0.703 ± 0.011
Llama3.2-Vision-11B (Meta, 2024)	37.77	30.99	0.849 ± 0.008	1.776 ± 0.010
KimiVL-16B (Team et al., 2025)	47.73	15.96	0.531 ± 0.005	0.788 ± 0.008
Kimi-k2.6 (Moonshot AI, 2026)	47.29	65.07	2.876 ± 0.009	6.383 ± 0.007
InternVL3.5-8B (Wang et al., 2025a)	43.77	19.74	1.415 ± 0.004	2.474 ± 0.013
GLM4.6V-9B (Z.ai, 2025)	44.51	37.78	1.302 ± 0.003	3.950 ± 0.007
Qwen3VL-8B (Bai et al., 2025a)	46.26	15.88	2.443 ± 0.011	1.704 ± 0.003
Qwen3.5-9B (Qwen Team, 2026a)	49.92	14.49	2.192 ± 0.006	3.324 ± 0.019
Qwen3.6-27B (Qwen Team, 2026b)	49.34	18.12	2.638 ± 0.010	4.364 ± 0.011
Gemma4-31B (Team et al., 2026)	50.22	34.31	2.819 ± 0.007	5.558 ± 0.013
Specialized Models				
Q-Instruct (Wu et al., 2024)	34.70	15.03	0.508 ± 0.003	0.170 ± 0.003
AesExpert (Huang et al., 2024a)	31.33	16.19	0.532 ± 0.005	0.031 ± 0.000
ArtiMuse (Cao et al., 2026)	42.75	13.18	0.977 ± 0.006	1.970 ± 0.007
CAPTURE	**70.28**	**75.64**	**5.340 ± 0.008**	6.860 ± 0.011



Figure 4: Qualitative retouching comparison. Each row shows the input, human expert ground truth (GT), and rendered outputs, with the target instruction below.

while the visual merger and language model remain trainable. We train for 3 epochs on 8 NVIDIA A6000 GPUs using FSDP2, with a global batch size of 32 and a maximum sequence length of 8192 tokens. We use AdamW with a learning rate of 1×10^{-5}, weight decay of 0.1, $\beta_1 = 0.9$, and $\beta_2 = 0.95$, together with gradient clipping at 1.0. The learning rate follows a cosine schedule with a linear warmup over the first 1% of training updates.

5.3 Quantitative Comparison on Photography Tasks

We compare CAPTURE with three proprietary and ten open-source general-purpose VLMs, with the evaluated model variants listed in Tabs. 2 and 3. We also evaluate against Q-Instruct (Wu et al., 2024), AesExpert (Huang et al., 2024a), and ArtiMuse (Cao et al., 2026) for low-level visual and aesthetic assessment, and JarvisArt (Lin et al., 2025), JarvisEvo (Lin et al., 2026), and MonetGPT (Dutt et al., 2025) for Lightroom parameter prediction. Tab. 2 and Tab. 3 report the results across perception, critique, and retouching. CAPTURE consistently outperforms competing methods across perception, critique, and retouching tasks. For perception, it exceeds the strongest baseline by 20.06 percentage points on attribute assessment. For critique, it matches GPT5.5 on single-image aes-

Table 3: VLM performance on retouching assistance. Recommendations use LLM-judge scores (mean ± standard deviation); Lightroom prediction uses PSNR/SSIM against expert edits.

Model	Editing Recommendation	Lightroom Parameter Prediction	
	Judge ↑	PSNR (dB) ↑	SSIM ↑
Proprietary Models			
GPT5.5	3.021 ± 0.021	16.06	0.6765
Gemini3.1-Pro	1.761 ± 0.006	18.27	0.7520
Qwen3.6-Plus	2.316 ± 0.008	16.95	0.7021
Open Source Models			
LLaVA-OneVision2-8B	1.382 ± 0.004	17.42	0.7805
Llama3.2-Vision-11B	1.768 ± 0.010	18.67	0.8000
KimiVL-16B	1.403 ± 0.024	17.69	0.7458
Kimi-k2.6	2.352 ± 0.013	16.33	0.6868
InternVL3.5-8B	1.413 ± 0.011	18.30	0.8023
GLM4.6V-9B	1.567 ± 0.007	18.20	0.7787
Qwen3VL-8B	1.690 ± 0.011	18.36	0.7839
Qwen3.5-9B	2.062 ± 0.006	17.77	0.7308
Qwen3.6-27B	2.019 ± 0.008	16.17	0.6850
Gemma4-31B	2.330 ± 0.017	17.12	0.7231
Specialized Models			
JarvisArt	/	14.17	0.5947
JarvisEvo	/	15.98	0.6841
MonetGPT	/	20.06	0.8313
CAPTURE	**4.533 ± 0.021**	**22.05**	**0.8879**

Table 4: PDP results across vision encoders. Accuracy evaluates 26-way classification of tonal and color changes; SRCC measures rank agreement between predicted and measured change magnitudes. PhotoEncoder denotes our adapted encoder, with the backbone specified in parentheses.

Encoder backbone	Attribute Classification Acc. (%) ↑	Magnitude Regression SRCC ↑
Original Encoders		
Llama3.2-Vision Encoder	24.34	0.385
Qwen2.5VL Encoder	25.40	0.351
Qwen3VL Encoder	26.69	0.376
Qwen3.5 Encoder	29.23	0.396
Qwen3.6 Encoder	37.14	0.372
Gemma4 Encoder	24.64	0.325
LLaVA-OneVision2 Encoder	26.74	0.386
MoonViT	30.70	0.421
GLM4.6V Encoder	39.45	0.406
InternViT-300M	45.26	0.409
Adapted Encoders		
PhotoEncoder (InternViT-300M)	**65.56**	**0.541**
PhotoEncoder (Qwen3VL)	**66.21**	**0.575**

thetic assessment while outperforming all specialized models. For retouching, it leads in editing recommendation and Lightroom parameter prediction, with a 1.99 dB PSNR gain over MonetGPT.

5.4 Visualized Comparison

Fig. 4 illustrates limitations of competing methods in balancing the requested brightening with tonal preservation. For example, the floral-crown portrait in the second row calls for a dreamy, high-key rendition. GPT5.5 and Gemini3.1-Pro brighten the image but produce a washed-out appearance with reduced tonal variation in the face and clothing. JarvisEvo leaves the portrait dark, while JarvisArt produces harsh contrast that departs from the requested soft appearance. CAPTURE achieves a brighter rendition closer to the expert reference while retaining facial tonal variation and texture in the white garment. More visual results are shown in Appendix C.3.

5.5 User Preference Study

To complement the automatic evaluation, we recruit 50 participants through Prolific (Prolific, 2024) to compare CAPTURE with five methods over 20 source images, yielding 1,000 ranking responses. As shown in Fig. 5, CAPTURE achieves the lowest mean rank of 3.005, compared with 3.226 for GPT5.5 and 3.275 for MonetGPT. It also ranks in the top three in 63.3% of evaluations, versus 55.4% for GPT5.5. The rank distribution further shows fewer low-ranking judgments, with mass shifting from ranks 4–5 toward ranks 2–3. Together, these results show that CAPTURE also performs favorably under human preference evaluation.

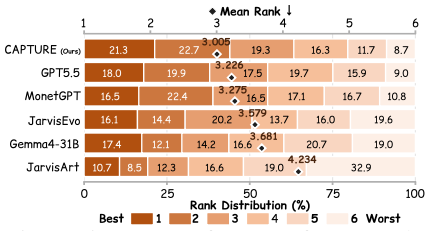


Figure 5: User preferences for retouching results. Stacked bars show the percentage of rankings. Black diamonds indicate mean ranks.

5.6 Effectiveness of PhotoEncoder

We evaluate ten original vision encoders, including Qwen2.5VL (Bai et al., 2025b), and two PhotoEncoder variants adapted from Qwen3VL and InternViT-300M with the PDP. Tab. 4 reports attribute classification accuracy and mean SRCC for magnitude regression. PhotoEncoder (Qwen3VL) achieves the highest performance on both tasks, reaching 66.21% classification accuracy and 0.575 SRCC. PhotoEncoder (InternViT-300M) similarly improves its corresponding backbone from 45.26% to 65.56% in classification accuracy and from 0.409 to 0.541 in SRCC. These consistent gains across two distinct architectures show that our adaptation generalizes beyond Qwen3VL and improves fine-grained photographic representations across vision backbones.

Table 5: Effect of PhotoEncoder under otherwise identical instruction-tuning settings. (a) Results on our benchmark, using the task names and metrics in Tabs. 2 and 3. (b) Pairwise win rates (%) on ArtiMuse-10K, evaluated by Qwen3.6-27B with expert analyses as references.

(a) Our benchmark

Model	Photographic Perception		Aesthetic Understanding		Retouching Assistance		
	Attribute Acc. (%) ↑	Transition Acc. (%) ↑	Aesthetic Critique Judge ↑	Style Comparison Judge ↑	Editing Recom. Judge ↑	LR Param. PSNR ↑	LR Param. SSIM ↑
CAPTURE	**70.28**	**75.64**	**5.340 ± 0.008**	**6.860 ± 0.011**	**4.533 ± 0.021**	**22.05**	**0.8879**
w/o PhotoEncoder	60.47	51.58	4.619 ± 0.006	6.071 ± 0.019	4.072 ± 0.023	20.94	0.8610

(b) ArtiMuse-10K: pairwise win rate (%) ↑

Model	Composition & Design	Visual Elements & Structure	Technical Execution	Originality & Creativity	Theme & Communication	Emotion & Viewer Response	Overall Gestalt	Comprehensive Evaluation	Overall
CAPTURE	**59.0**	**56.0**	**55.6**	**53.2**	**54.1**	49.5	**54.6**	**51.9**	**54.2**
w/o PhotoEncoder	41.0	44.0	44.4	46.8	45.9	**50.5**	45.4	48.1	45.8

Table 6: Ablations of the losses in PhotoEncoder adaptation. The PDP evaluates tonal and color representation; retrieval evaluates content caption alignment.

ID	$\mathcal{L}_{\text{attr}}$	$\mathcal{L}_{\text{edit}}$	$\mathcal{L}_{\text{content}}$	$\mathcal{L}_{\text{anchor}}$	Attribute Classification Acc. (%) ↑	Magnitude Regression SRCC ↑	Image-to-Text R@1 (%) ↑	Text-to-Image R@1 (%) ↑
#1	✓	✓	✓	✓	66.21	0.575	94.03	96.07
#2	✗	✓	✓	✓	64.78	0.538	94.96	97.20
#3	✓	✗	✓	✓	35.57	0.477	92.16	91.78
#4	✓	✓	✗	✓	67.61	0.611	0.75	6.73
#5	✓	✓	✓	✗	66.37	0.594	92.16	94.39

5.7 Ablation Studies

PhotoEncoder. To isolate the contribution of PhotoEncoder, we replace it with the original Qwen3VL vision encoder and repeat photographic instruction tuning using identical data and training settings. As shown in Tab. 5(a), PhotoEncoder improves all seven downstream metrics. We further evaluate whether these gains extend to fine-grained aesthetic analysis on ArtiMuse-10K. Following the ArtiMuse (Cao et al., 2026) evaluation dimensions, we conduct pairwise comparisons with Qwen3.6-27B as the judge to select the better analysis between the two model variants. As shown in Tab. 5(b), the PhotoEncoder variant achieves an overall win rate of 54.2% versus 45.8% for the original encoder, outperforming it in seven of the eight evaluation dimensions. These results demonstrate that the benefits of representation-level adaptation generalize beyond our benchmark, yielding consistent improvements on an external aesthetic dataset.

Loss Components. We ablate each objective in Tab. 6 to examine the trade-off between photographic sensitivity, measured by the PDP, and preservation of pretrained semantics, measured by image–text retrieval. Relative to the full objective (#1), removing $\mathcal{L}_{\text{attr}}$ (#2) improves retrieval slightly (94.96/97.20 vs. 94.03/96.07 R@1), but weakens both photographic probes, reducing attribute classification from 66.21% to 64.78% and magnitude SRCC from 0.575 to 0.538. This trade-off is undesirable because improving fine-grained photographic representation is the primary objective of PhotoEncoder adaptation. Conversely, removing $\mathcal{L}_{\text{anchor}}$ (#5) yields modestly higher probe scores (66.37% and 0.594), but causes a larger drop in retrieval to 92.16%/94.39%, indicating increased drift from the pretrained semantic representation. The full objective (#1) therefore provides a better balance, retaining most of the probe gains from representation adaptation while preserving the encoder's original semantic capability. Rows #3 and #4 further support the distinct roles of the objectives. Removing $\mathcal{L}_{\text{edit}}$ (#3) severely degrades attribute discrimination, whereas removing $\mathcal{L}_{\text{content}}$ (#4) largely destroys image–text retrieval despite improving the probe metrics.

6 Conclusion

We present CAPTURE, a framework for photographic assistance. Using the PDP, we identify limited representation of fine-grained tonal and color changes in existing vision encoders, motivating representation-level adaptation. We introduce PhotoEncoder to capture photographic attributes and editing-induced changes while preserving pretrained representations, and integrate it with an LLM through photographic instruction tuning. Experiments show that CAPTURE consistently improves downstream photographic tasks, demonstrating gains from representation-level adaptation.

AI USE STATEMENT

We used generative AI primarily to improve the manuscript's language and clarity, and also to assist with drafting text and writing code. We used VLMs to generate synthetic textual supervision from source images, expert-retouched images, and editing records, including aesthetic critiques, style analyses, and editing recommendations, as described in Section 5.1. Models also assisted in generating benchmark reference answers and scoring rubrics, and Qwen3.6-27B served as the judge for open-ended responses. The authors take responsibility for the final manuscript, implementation, data, and reported results, including content produced with AI assistance.

ETHICS STATEMENT

This work is intended to support user-directed photographic critique and retouching. The data are derived from MIT-Adobe FiveK and PPR10K, and aesthetic supervision reflects the preferences of their expert editors and the models used for annotation and evaluation. Such judgments may carry stylistic or cultural biases and should not be treated as universal standards of photographic quality. Users should retain control over editing decisions, particularly when changes affect how people or scenes are represented. Reuse and redistribution of source photographs and derived data should respect the source datasets' licenses, image rights, and the privacy of depicted individuals.

REPRODUCIBILITY STATEMENT

We provide additional technical details in the supplementary material and will release our code and datasets to support reproducibility.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Fatemeh Behrad, Tinne Tuytelaars, and Johan Wagemans. Charm: the missing piece in vit fine-tuning for image aesthetic assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7815–7824. IEEE, 2025.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 97–104. IEEE, 2011.

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, et al. Unipercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. *arXiv preprint arXiv:2512.21675*, 2025.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 15313–15322, 2026.

Haoyu Chen, Keda Tao, Yizao Wang, Xinlei Wang, Lei Zhu, and Jinjin Gu. Photoartagent: Intelligent photo retouching with language model-based artist agents. *arXiv preprint arXiv:2505.23130*, 2025a.

Xiangyu Chen, Zheyuan Li, Zhengwen Zhang, Jimmy S Ren, Yihao Liu, Jingwen He, Yu Qiao, Jiantao Zhou, and Chao Dong. Towards efficient sdrtv-to-hdrtv by learning from image formation. *IEEE Trans. Multimedia*, 2025b.

Jinguang Cheng, Chunxiao Li, Shuai He, Taiyu Chen, and Anlong Ming. Thinking aesthetics assessment of image color temperature: models, datasets and benchmarks. In *Proc. AAAI Conf. Artif. Intell.*, volume 40, pp. 3246–3254, 2026.

Ming-Chang Chiu, Shicheng Wen, Pin-Yu Chen, and Xuezhe Ma. Megacoin: Enhancing medium-grained color perception for vision-language models. In *Proc. AAAI Conf. Artif. Intell.*, volume 40, pp. 37305–37313, 2026.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Venus: Benchmarking and empowering multimodal large language models for aesthetic guidance and cropping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 37766–37776, June 2026.

Niladri Shekhar Dutt, Duygu Ceylan, and Niloy J Mitra. Monetgpt: Solving puzzles enhances mllms' image retouching skills. *ACM Trans. Graph.*, 44(4):1–12, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *Eur. Conf. Comput. Vis.*, pp. 148–166. Springer, 2024.

Google. Gemini 3.1 pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, February 2026.

Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, and Huadong Ma. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *Int. Conf. Comput. Vis.*, pp. 21781–21790. IEEE, 2023a.

Shuai He, Anlong Ming, Shuntian Zheng, Haobin Zhong, and Huadong Ma. Eat: An enhancer for aesthetics-oriented transformers. In *ACM Int. Conf. Multimedia*, pp. 1023–1032, 2023b.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Int. Conf. Learn. Represent.*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

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 Int. Conf. Multimedia*, pp. 5911–5920, 2024a.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024b.

Ziheng Jia, Linhan Cao, Jinliang Han, Zicheng Zhang, Jiaying Qian, Jiarui Wang, Zijian Chen, Guangtao Zhai, and Xiongkuo Min. Vital: Vision-encoder-centered pre-training for lmms in visual quality assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 41245–41255, 2026.

Hyeongmin Lee, Kyoungkook Kang, Jungseul Ok, and Sunghyun Cho. Cliptone: Unsupervised learning for text-based image tone adjustment. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2942–2951. IEEE, 2024.

Changjun Li, Zhiqiang Li, Zhifeng Wang, Yang Xu, Ming Ronnier Luo, Guihua Cui, Manuel Melgosa, Michael H Brill, and Michael Pointer. Comprehensive color solutions: Cam16, cat16, and cam16-ucs. *Color Research & Application*, 42(6):703–718, 2017.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Jian Zhang, et al. Q-insight: Understanding image quality via visual reinforcement learning. In *Adv. Neural Inform. Process. Syst.*, volume 38, pp. 36802–36827, 2025.

Zhendong Li, Lei Sun, Ruibo Ming, He Zhang, Danda Pani Paudel, Luc Van Gool, and Jinjin Gu. Knowledge-centric agents for workflow generation in comfyui. In *Eur. Conf. Comput. Vis.*, pp. 306–323. Springer, 2026.

Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 653–661. IEEE, 2021.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. In *Adv. Neural Inform. Process. Syst.*, volume 38, 2025.

Yunlong Lin, Zixu Lin, Kunjie Lin, Jinbin Bai, Panwang Pan, Chenxin Li, Haoyu Chen, Zhongdao Wang, Xinghao Ding, Wenbo Li, et al. Jarvisart: Liberating human artistic creativity via an intelligent photo retouching agent. In *Adv. Neural Inform. Process. Syst.*, volume 38, pp. 52088–52130, 2025.

Yunlong Lin, Linqing Wang, Kunjie Lin, Zixu Lin, Kaixiong Gong, Wenbo Li, Bin Lin, Zhenxi Li, Shiyi Zhang, Yuyang Peng, et al. Jarvisevo: Towards a self-evolving photo editing agent with synergistic editor-evaluator optimization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 27291–27302, 2026.

M Ronnier Luo, Guihua Cui, and Bryan Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(5):340–350, 2001.

Meta. Llama 3.2 Vision model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD_VISION.md`, September 2024.

Moonshot AI. Kimi K2.6: Advancing open-source coding. `https://www.kimi.ai/blog/kimi-k2-6`, April 2026.

OpenAI. Gpt-5.5 system card. `https://openai.com/index/gpt-5-5-system-card/`, April 2026.

Prolific. Prolific - participant recruitment for research. `https://www.prolific.com`, 2024. Accessed: 2025-05-24.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 24807–24816. IEEE, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026a.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model. `https://qwen.ai/blog?id=qwen3.6-27b`, April 2026b.

Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026c. URL `https://qwen.ai/blog?id=qwen3.6`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pp. 8748–8763. PmLR, 2021.

Dylan Sam, Devin Willmott, Joao D Semedo, and J Zico Kolter. Finetuning clip to reason about pairwise differences. *arXiv preprint arXiv:2409.09721*, 2024.

Alvy Ray Smith. Color gamut transform pairs. *ACM Siggraph Computer Graphics*, 12(3):12–19, 1978.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9568–9578. IEEE, 2024.

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

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proc. AAAI Conf. Artif. Intell.*, volume 37, pp. 2555–2563, 2023.

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*, 2025a.

Zhihua Wang, Yu Long, Qiuping Jiang, Chao Huang, and Xiaochun Cao. Harnessing multi-modal large language models for measuring and interpreting color differences. *IEEE Trans. Image Process.*, 2025b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 25490–25500. IEEE, 2024.

Qiucheng Wu, Jing Shi, Simon Jenni, Kushal Kafle, Tianyu Wang, Shiyu Chang, and Handong Zhao. Retouchiq: Mllm agents for instruction-based image retouching with generalist reward. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 12279–12288, June 2026.

Mingde Yao, Zhiyuan You, King-Man Tam, Menglu Wang, and Tianfan Xue. Photoagent: Agentic photo editing with exploratory visual aesthetic planning. *arXiv e-prints*, pp. arXiv–2602, 2026.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *Eur. Conf. Comput. Vis.*, pp. 259–276. Springer, 2024.

Zhiyuan You, Jinjin Gu, Xin Cai, Zheyuan Li, Kaiwen Zhu, Chao Dong, and Tianfan Xue. Enhancing descriptive image quality assessment with a large-scale multi-modal dataset. *IEEE Trans. Image Process.*, 34:8201–8215, 2025.

Z.ai. GLM-4.6V-Flash model card. `https://huggingface.co/zai-org/GLM-4.6V-Flash`, 2025.

Ying Zeng, Miaosen Luo, Guangyuan Li, Yang Yang, Ruiyang Fan, Linxiao Shi, Qirui Yang, Jian Zhang, Chengcheng Liu, Siming Zheng, et al. Smartphotocrafter: Unified reasoning, generation and optimization for automatic photographic image editing. *arXiv preprint arXiv:2604.19587*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Int. Conf. Comput. Vis.*, pp. 11941–11952. IEEE, 2023.

Shuo Zhang and Xinyu Yang. Retouchagent: Towards interactive and explainable image retouching with mllm agents. In *Proc. AAAI Conf. Artif. Intell.*, volume 40, pp. 29901–29910, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Adv. Neural Inform. Process. Syst.*, volume 36, pp. 46595–46623, 2023.

Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024.

Zichen Zhu, Yuheng Sun, Mingxuan Zhu, Wenjie Ma, Situo Zhang, Zhexiang Wang, Ziyue Yang, Danyang Zhang, Kunyao Lan, Zihan Zhao, et al. Iea: Amateur-friendly conversational image editing agent via three stages of multitask alignment. *arXiv preprint arXiv:2606.08016*, 2026.