PhotoFramer: Multi-modal Image Composition Instruction

Zhiyuan You[12], Ke Wang[3], He Zhang[4], Xin Cai[2], Jinjin Gu[5],
Tianfan Xue[267†], Chao Dong[168†], Zhoutong Zhang[3]
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]Multimedia Laboratory, The Chinese University of Hong Kong
[3]Adobe NextCam [4]Adobe Research [5]INSAIT, Sofia University "St. Kliment Ohridski"
[6]Shanghai AI Laboratory [7]CPII under InnoHK [8]Shenzhen University of Advanced Technology
zhiyuanyou@foxmail.com, tfxue@ie.cuhk.edu.hk, chao.dong@siat.ac.cn, zhoutongz@adobe.com
Project Page: https://zhiyuanyou.github.io/photoframer [†] Corresponding Author

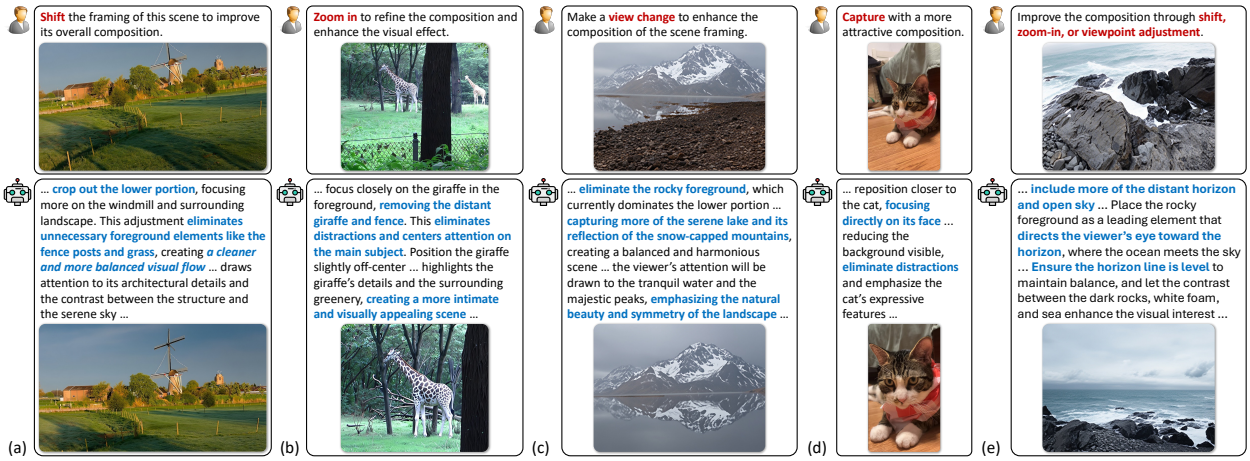

Figure 1. We propose PhotoFramer, a model designed for composition instruction during photo capturing. Given a poorly composed image, PhotoFramer first describes how to improve the composition in natural language and then generates an example image that follows the described suggestions. The photo-taker can follow the textual guidance and the example image to capture a better-composed photo.

Abstract

Composition matters during the photo-taking process, yet many casual users struggle to frame well-composed images. To provide effective composition guidance, we introduce PhotoFramer, a multi-modal composition instruction framework. Given a poorly composed image, PhotoFramer first describes how to improve the composition in natural language and then generates a well-composed example image. To train such a model, we curate a large-scale dataset. Inspired by how humans take photos, we organize composition guidance into a hierarchy of sub-tasks: shift, zoom-in, and view-change tasks. Shift and zoom-in data are sampled from existing cropping datasets, while view-change data are obtained via a two-stage pipeline. First, we sample pairs with varying viewpoints from multi-view datasets, and train a degradation model to transform well-composed photos into poorly composed ones. Second, we apply this degradation model to expert-taken photos to synthesize poor images to form training pairs. Using this dataset, we finetune a model that jointly processes and generates both text and images, enabling actionable textual guidance with illustrative examples. Extensive experiments demonstrate that textual instructions effectively steer image composition, and coupling them with exemplars yields consistent improvements over exemplar-only baselines. PhotoFramer offers a practical step toward composition assistants that make expert photographic priors accessible to everyday users.

1. Introduction

Modern mobile cameras have become increasingly powerful [8, 20, 39, 52]. Equipped with these cameras, the users could capture high-resolution, noise-free, and well-exposed photos. However, many casual users still struggle to capture visually pleasing photos. As shown in the top images of Fig. 1, these photos appear visually unappealing due to their poor composition. For instance, in Fig. 1e, the sea horizon is tilted relative to the image border, and the foreground

rocks dominate the frame excessively. Therefore, providing amateur photographers with composition guidance during photo capturing is of great importance [13, 31, 73].

Casual users can be guided more effectively through a combination of textual and visual instructions. Consider a captured farm scene in Fig. 1a, textual instruction "eliminates the fence posts and grass" provides concrete and actionable operations, along with detailed reasons (*e.g.*, "creating a cleaner and more balanced visual flow"). Meanwhile, in the bottom of Fig. 1a, the well-composed example photo capturing the same farm scene is intuitive and easy to follow, consistent with prior findings [13, 76].

To leverage the strengths of both modalities, we introduce PhotoFramer, a multi-modal image composition instruction framework that provides detailed textual guidance paired with corresponding visual example photos.

To build this composition instruction model, we design a hierarchical set of tasks inspired by how humans take photos. Humans typically first determine a suitable position and angle (*i.e.*, vantage point), choose an appropriate focal length, and then finetune subject placement and alignment [14, 26]. Accordingly, PhotoFramer consists of three sub-tasks: view-change (Fig. 1c), zoom-in (Fig. 1b), and shift (Fig. 1a) tasks. In addition, when the user does not specify a task type (as in Fig. 1de), PhotoFramer automatically determines suitable operations. We show that the model does not merely select one task type, but adaptively fuses multiple operations to achieve better composition.

Under this task formulation, we construct a comprehensive multi-modal dataset containing "poor image, good image, text guidance" triplets. For the shift and zoom-in sub-tasks, we sample image pairs from existing cropping datasets [6, 58, 65, 66, 72]. For the view-change sub-task, we first sample image pairs with different viewpoints from multi-view datasets [34], and then train a degradation model that converts well-composed photos into compositionally degraded ones. We then apply this model to expert-taken photos [10] to synthesize view-change examples. Finally, we employ one vision-language model [2] to annotate text guidance for each pair. In total, we collect 207K triplets, serving as the foundation for model training.

Equipped with the above dataset, we finetune the unified understanding-generation model [9, 16, 51, 62, 64] to generate both textual guidance and example image. As depicted in Fig. 1, our model is required to process inputs and outputs that include both texts and images. Consequently, vision-language models [2, 37, 78] that can only produce textual outputs, as well as image generation or editing models [3, 4, 47] that can only produce images, are unsuitable. We therefore employ the unified understanding-generation model Bagel [9] as our base and fine-tune it using our proposed dataset. Experiments show that textual guidance effectively guides image generation, highlighting the advantages of the text-vision joint modeling framework.

Extensive comparisons and ablations further confirm the superiority of our PhotoFramer over baselines. We hope this work serves as a stepping stone toward composition assistants that help everyday users capture expert-level photos.

2. Related Works

Composition understanding is the foundation for composition instruction. Composition classification [27, 73, 77] defines multiple composition categories (*e.g.*, rule-of-thirds, symmetrical, *etc.*). Some works [71–73] focus on score-based composition assessment. Recent works [5, 35] utilize powerful vision language models [2, 41, 79] to build multi-aspect aesthetics (*i.e.*, including composition) assessment models with joint scoring and descriptive outputs.

Image cropping is a common approach to enhance the composition. Given an original image, this task aims to find a cropped patch with better composition. Extensive cropping datasets can be broadly categorized into two types. The first type is densely annotated, where multiple crops are labeled per image [49, 58, 66, 71, 72]. The second type is sparsely annotated, where only the best crop is labeled per image [6, 11, 65]. Built upon these datasets, deep learning models have achieved substantial progress. Many methods [11, 22, 30, 49, 55, 57, 58, 65, 66, 71, 72, 74] adopt a two-stage strategy: first generating crop candidates, then selecting the optimal one. Other coordinate regression methods [7, 17, 18, 28, 29, 38] directly predict crop boundaries. Recently, GenCrop [19] leverages Stable Diffusion [47] to outpaint professional photos for dataset expansion. ProCrop [75] retrieves compositionally similar reference images to guide the cropping process. However, image cropping remains a post-processing operation applied after photo capture. In contrast, our work aims to guide users *during* the capturing process to take well-composed photos.

Composition guidance. Traditional methods rely on retrieval to search similar images in the database as user reference [12, 13, 76]. However, retrieval-based guidance suffers from scene and subject mismatch, making it hard to follow. Recent CPAM [31] could automatically provide photographers with camera pose adjustment guidance. Our work differs from CPAM in three key aspects. First, CPAM predicts only adjusted yaw and pitch angles, while we generate both text instruction and a good example image, offering a more intuitive and informative guide. Second, CPAM is limited to yaw and pitch adjustments, whereas our model additionally supports zoom control and large-scale viewpoint changes. Third, CPAM employs separate models for understanding and adjustment, while we adopt a unified model that jointly performs both tasks, enabling mutual enhancement.

Unified multi-modal model conducts image understanding and generation in a unified model [9, 16, 50, 51, 53, 59, 62, 64]. Since our aim is to output both textual instructions

(*i.e.*, understanding) and example images (*i.e.*, generation), we take the unified multi-modal model as our base model.

3. Task Paradigm and Dataset Construction

3.1. Task Paradigm

We first construct the task paradigm for composition guidance. Our goal is to develop an assistant that can guide humans during the photo-taking process. Therefore, we begin by revisiting how humans take photos and analyzing the key abilities the model should possess. We summarize three key steps to decompose the capturing process from [15].

– First, humans choose an appropriate shooting position and angle, referred to as the *vantage point*. As shown in Fig. 2c, given a scene, our model needs to infer alternative viewpoints and select the one with the best composition.

– Second, humans adjust the focal length (or zoom level on mobile photo cameras) to emphasize specific subjects. Accordingly, PhotoFramer should have the ability to identify well-composed crops within a larger scene (Fig. 2b).

– Third, humans carefully adjust the camera to maintain a level frame, avoid border distractions, and place subjects in balanced positions. Thus, as shown in Fig. 2a, PhotoFramer should position subjects appropriately (*e.g.*, centered or rule of thirds) and remove distractions to maintain clean borders.

Based on the above discussion, as illustrated in Fig. 2, we design a hierarchical task paradigm that progressively guides our PhotoFramer to acquire these capabilities:

– *Shift task*. Given a poorly composed image, PhotoFramer adjusts the framing to properly place the subject, levels the image, and removes border distractions.

– *Zoom-in task*. Given an original image, PhotoFramer generates a tighter crop with improved composition.

– *View-change task*. Given a captured scene, PhotoFramer selects a new vantage point or camera pose to reframe the scene and generates the corresponding image.

Moreover, both textual guidance and example images are important. Example images are intuitive and easy to follow, while the textual guidance provides detailed reasoning for the generated image and is easier to understand. Therefore, as depicted in Fig. 2, our PhotoFramer is designed to generate a detailed *text guidance* (describing how to improve the composition) together with an *example image* (demonstrating what a well-composed image should look like).

Formally, let the input poor-composition image be denoted as `I_poor`, the task type (*i.e.*, expressed in text) as `T_task`, the generated text guidance as `T_guide`, and the target well-composed image as `I_good`. Our PhotoFramer, denoted as `f()`, is trained to perform the following mapping: `I_good, T_guide = f(I_poor, T_task)`.

3.2. Dataset Construction

Data is the key factor in training such a unified model. Following [9, 53], we need to construct `<T_task, I_poor, I_good, T_guide>` pairs for supervision.

Task prompt collection. For `T_task`, we follow [61, 67, 68] by predefining some text templates for each task and randomly sampling one template to form the pair. For example, a template for the shift task is "shift the scene to enhance the composition". See Appendix for all task prompts.

Image pair collection. Collecting `<I_poor, I_good>` image pairs is the key process and will be detailed later.

Text guidance collection. Given `<I_poor, I_good>` image pairs, we employ a vision-language model, Qwen2.5-VL-32B [2], to generate text guidance `T_guide`. Specifically, we input the poor and good images along with the task type, and prompt the model to describe how to transform the poor image into the good one, with detailed justifications.

Dataset statistics are given in Tab. 1. Since the shift and zoom-in pairs are sampled from the original images in the cropping datasets (see follows), both the original images and pairs are included in the statistics. In total, our constructed dataset comprises 45K original images and 207K pairs, providing a solid foundation for model training.

Table 1. Dataset statistics.

	Shift	Zoom-in	View-change
# Original	10,321	7,665	27,393
# Pairs	164,904	14,182	

3.2.1. Shift and Zoom-in Pairs

As illustrated in Fig. 3, we construct shift and zoom-in pairs from existing cropping datasets, which offers two advantages. First, the scenes and subjects in cropping datasets are curated by contributors, making them well-suited for composition-related training. Second, they contain useful annotations, which could reduce annotation workload.

Shift pair collection. We collect shift pairs from existing cropping datasets GAIC [72] and CPC [58]. GAIC provides scores for each crop, while CPC contains only selected good and best crops, which cannot be directly used. To generate missing scores in CPC, we design a mathematical sampling model in Appendix. As shown in Fig. 3, crops with scores above 4.0[1] are treated as good images, while those below 2.0 are treated as poor images, forming `<I_poor, I_good>` pairs. A small proportion of mid-score crops are sampled as poor images to enhance robustness. Finally, a random rotation is applied to the poor crop for augmentation. Note that our work focuses on composition instructions without changing aspect ratio. Hence, we discretize the aspect ratio range [0.45, 2.2] into 11 values, and only images with the same aspect ratio could be paired.

Zoom-in pair collection. We pair each good crop with its original image to form a `<I_poor, I_good>`[2] pair, as depicted in Fig. 3. Thus, the key step is to identify the good crop. We adopt existing cropping datasets including GAIC [72], CPC [58], SACD [66], FlickrCrop [6], FLMS [11], and CUHKCrop [65]. For GAIC and CPC,

[1]Composition scores in this work are all normalized to a [1,5] scale.

[2]We use "poor" for simplicity. The original image may not be strictly poor.

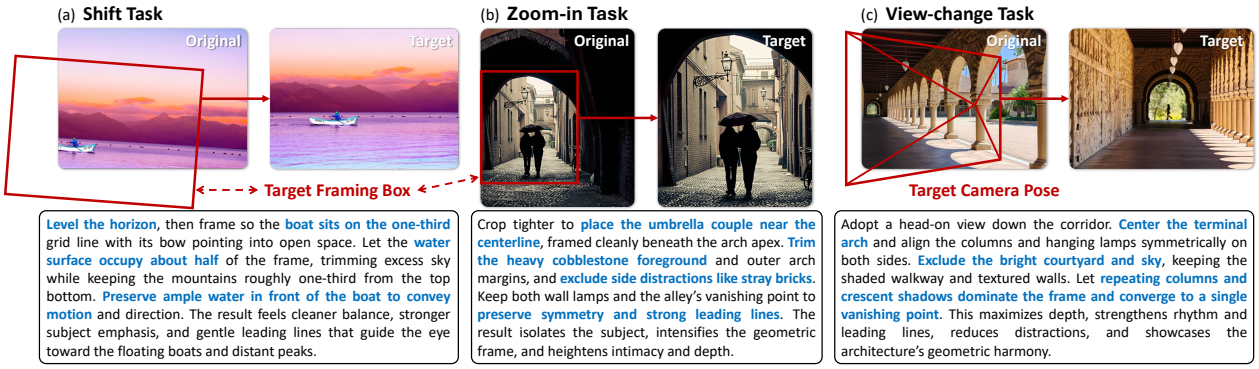

Figure 2. Task paradigm and data example. Given a poorly composed image, our PhotoFramer is required to generate a *textual guidance* (describing how to improve the composition) together with an *example image* (depicting what a well-composed image looks like). Motivated by three key photography factors (*vantage point*, *focal choice*, and *subject placement*), our PhotoFramer comprises three tasks: (a) **Shift**: adjust the framing to place the subject properly and remove border distractions; (b) **Zoom-in**: select a tighter crop (simulating a longer focal length) that yields a stronger composition; (c) **View-change**: choose a new vantage point or camera pose to reframe the scene.

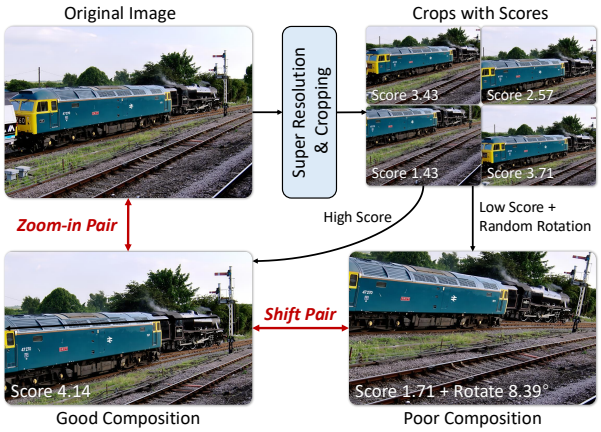

Figure 3. Dataset construction for the shift and zoom-in tasks. For the shift task, given an image from the cropping dataset, we sample its crops to form a `<poor,good>` image pair. A random rotation is applied to the poor crop. For the zoom-in task, the original image and a well-composed crop form an `<original,good>` pair. To ensure sufficient resolution, we apply $4\times$ super resolution to the original image using HYPIR [33].

where a score can be assigned to each crop, crops with scores above 4.0 are regarded as good crops. Other datasets provide a human-labeled best crop, which can be directly used. To ensure the two paired images share the same aspect ratio, we crop the original image with the largest possible region that matches the good crop's aspect ratio. The crop center is chosen as close as possible to the image center, subject to fully containing the good crop.

Super-resolution for sufficient resolution. Some crops have very low resolution, *i.e.*, below 300 pixels, which is hard to use. Therefore, before cropping, we apply a $4\times$ super-resolution to the original image using HYPIR [33].

Image pair filtering. Three types of filtering are performed to ensure data quality. (1) Pairs containing any image with an aspect ratio outside [0.45, 2.2] are discarded. (2) For shift pairs, we remove pairs depicting different subjects. (a) Pairs with CLIP similarity [46] below 0.8 are discarded. (b) We use U^2Net [45] to extract saliency masks of the subjects, compute DINOv2 [44] features on the masked regions, and discard pairs with subject-level cosine similarity below 0.6. (c) To avoid confusion with the zoom-in task, we filter out pairs where one image is fully contained in the other and the area ratio is below 0.6. (d) To ensure sufficient composition difference in each pair, we retain those pairs whose good score exceeds the poor score by at least 0.8. (3) For zoom-in pairs, to avoid trivial cases where the good crop nearly overlaps with the original, we filter out pairs in which the good crop occupies more than 60% of the original image.

Table 2. Results of our assessment model. Metrics for assessment and classification tasks are SRCC / PLCC and accuracy. CADB [73] and GAIC [72] datasets are for composition assessment, while AVA [40] dataset is for aesthetic assessment.

Task	Composition / Aesthetic Assessment			Classif.
Dataset	CADB	GAIC	AVA	CADB
VFN [7]	0.052 / 0.049	0.152 / 0.162	0.139 / 0.142	-
VEN [58]	0.084 / 0.082	0.410 / 0.428	0.232 / 0.241	-
AutoPhoto [1]	0.065 / 0.079	0.407 / 0.427	0.604 / 0.613	-
Q-Align [61]	0.561 / 0.557	0.169 / 0.178	0.809 / 0.804	-
Qwen2.5-VL-32B [2]	0.420 / 0.426	0.195 / 0.205	0.527 / 0.492	0.101
Our Model (7B)	**0.763 / 0.777**	**0.795 / 0.805**	**0.825 / 0.828**	**0.583**

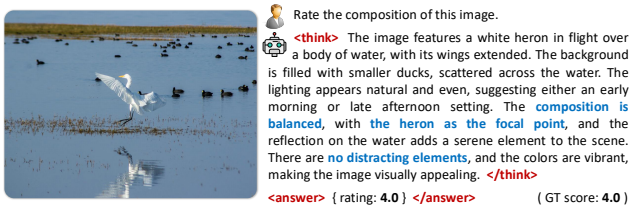

Figure 4. Qualitative results of our composition assessment model, illustrating the thinking process and final assessment output.

3.2.2. Composition Assessment Model

To construct the remaining view-change pairs, a good composition assessment model is necessary. As described in

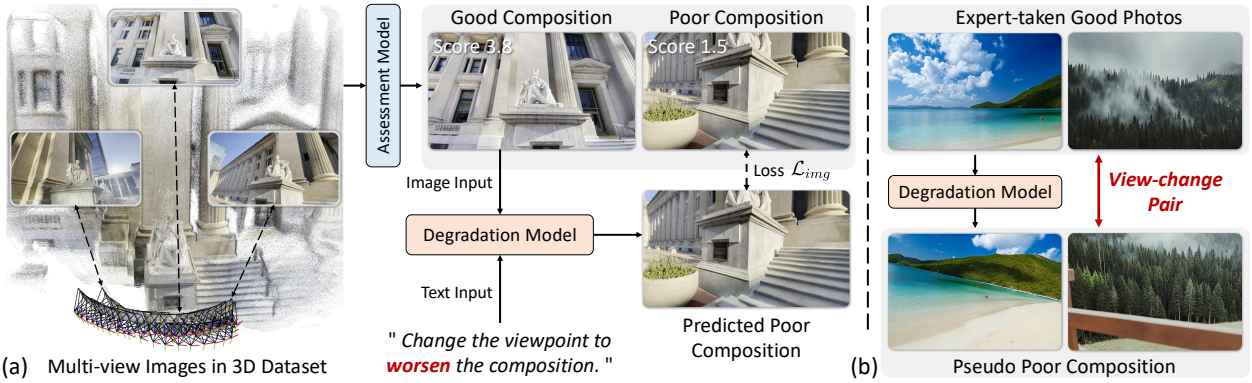

Figure 5. Dataset construction for the view-change task. (a) Leveraging our composition assessment model in Sec. 3.2.2, we sample <poor,good> image pairs from multi-view datasets. We then train a composition degradation model that generates poor-composition images from good ones. (b) We apply this degradation model to expert-taken good photos to synthesize pseudo poor-composition images, forming the final pairs. We do not rely solely on multi-view datasets, as most of their good images are not sufficiently well-composed.

Sec. 3.2.1, the shift and zoom-in pairs are built upon cropping datasets with human-provided annotations. However, most in-the-wild images lack such annotations.

Composition assessment dataset. To train such an assessment model, we primarily use composition scoring datasets, CADB [73] and GAIC [72], and additionally incorporate composition classification datasets, CADB [73] and KU-PCP [27], and the aesthetic assessment dataset AVA [40], given their strong relevance to composition assessment.

Composition assessment model. Following [32, 63], we adopt Qwen2.5-VL-7B [2] as the base model and train it using the GRPO [48] reinforcement learning algorithm. The composition assessment and classification results are presented in Tab. 2. Our 7B model outperforms all prior assessment models [1, 7, 58], Q-Align [61], and the much larger Qwen2.5-VL-32B [2]. One qualitative example is shown in Fig. 4, where our model provides detailed reasoning texts alongside the final accurate composition score.

3.2.3. View-change Pairs

Pair collection from multi-view data. A natural source to sample image pairs with varying viewpoints is 3D datasets containing multi-view images. DL3DV-10K [34] is a large-scale 3D dataset comprising 10K scenes and 51M frames. As shown in Fig. 5, we evaluate images within each scene using our assessment model, then select up to three best images and ten worst images to form pairs. However, even the best frames often lack expert-level composition quality, making it insufficient to rely solely on this dataset.

Pair collection from expert-taken photos. Multi-view data can provide image pairs, but the good images often lack strong composition quality. In contrast, expert-taken photos exhibit good composition but lack corresponding poor images to form pairs. Thus, as shown in Fig. 5, we use the expert-taken photo as good image, and generate a poor view of this image, forming the view-change pair.

– *Composition degradation model*. In the first stage, we use the pairs from 3D dataset to train a degradation model (the same to the final model introduced in Sec. 4). As depicted in Fig. 5a, the model takes a well-composed image along with a degradation instruction such as "Change the viewpoint to *worsen* the composition", and generates the corresponding poor-composition image. An image reconstruction loss between the predicted poor image and the ground-truth poor image is minimized to optimize the model.

– *Degradation on excellent images.* As shown in Fig. 5b, in the second stage, we apply the trained degradation model to human-taken photos to synthesize pseudo poor images, forming the final pairs. We use two sources of good images: the Unsplash Lite dataset [10] with 25K professional photographs, and another 10K images taken by ourselves. The Unsplash Lite dataset reflects professional-level photography, while our own dataset represents the level of amateur photography, providing complementary data diversity.

Image pair filtering. We perform three types of filtering to ensure data quality. (1) The first discards pairs containing any image with an extreme aspect ratio outside [0.45, 2.2]. (2) The second aims to ensure the quality of good images. (a) File size: we discard images smaller than 200 KB (JPG, 1024 resolution, quality 95), as such images are often overly simplistic (*e.g.*, plain colors or curves). (b) Image quality: images with a DeQA-Score [69] below 3.5 are removed. (c) Composition: only images with a composition score above 3.0 are retained. (d) Art: for the Unsplash Lite dataset [10], we filter out images with keywords like "abstract", "painting", and "art", as our focus is on real-world scenes. (3) The third aims to enforce view consistency within each image pair. VGGT [56] is employed to compute the Field-of-View (FoV) overlap between pairs. Low FoV intersection means poor spatial correspondence and thus leads to exclusion.

4. PhotoFramer Model

Model architecture. As stated in Sec. 3.1, our goal is to predict both a textual instruction and a well-composed ex-

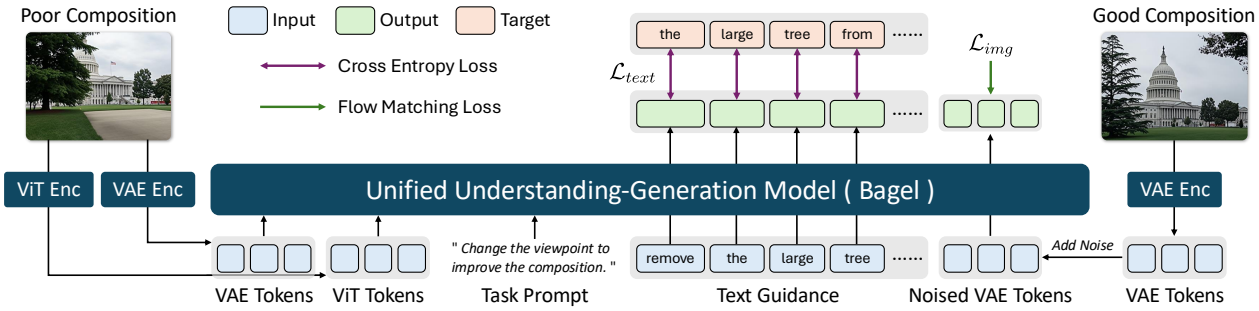

Figure 6. PhotoFramer architecture. We adopt Bagel [9], a unified understanding–generation model, as the base model. Given a task prompt and a poorly composed image, the model predicts both text guidance and a well-composed example image. The text guidance is optimized using a cross-entropy loss $\mathcal{L}_{text}$ for next-token prediction, while the example image is trained using a flow matching loss $\mathcal{L}_{img}$.

ample image. Bagel [9] is a unified multi-modal model capable of producing both textual and visual outputs, with strong language understanding, visual perception, and image generation capabilities. Therefore, we adopt Bagel as our base model for finetuning. As illustrated in Fig. 6, given a task prompt and a poorly composed image, the model predicts a textual instruction and a well-composed example image. For visual input, Bagel uses two types of vision tokens: VAE tokens encoded by the FLUX VAE [25] and ViT tokens extracted by the SigLIP2-so400m/14 [54] ViT model. The VAE tokens, containing pixel-level information, are used for image generation, while the ViT tokens, encoding semantic information, are leveraged for visual understanding. The textual instruction is optimized with a cross-entropy loss $\mathcal{L}_{text}$ for next-token prediction, and the example image is optimized with a flow-matching loss $\mathcal{L}_{img}$. Note that the generated image attends to the textual instruction through the attention mechanism, enabling instruction-driven example image generation. For further architectural details of Bagel, we refer the reader to [9].

Auto prompt design. As shown in Fig. 2, PhotoFramer supports three sub-tasks. Each task uses predefined prompt templates. However, this requires users to explicitly specify the task type, which can be inconvenient. To alleviate this issue, we adopt *auto prompts* such that the user does not need to indicate the task. We propose two types of auto prompts: (1) *Static auto task*, where users may not want to change their viewpoint, thus only shift and zoom-in are allowed. (2) *Full auto task*, where all three tasks could be applied. For static auto prompts, we use, *e.g.*, "Refine the composition through shift or zoom-in adjustments". For full auto prompts, we use free-form instructions such as "Capture this scene with better composition". More examples are provided in Appendix. During training, task-specific prompts are randomly replaced by auto prompts, enabling the model to self-determine the most suitable operations.

Inference. The inference process consists of two stages. (1) *Understanding*. The model takes visual tokens and a task prompt as inputs, analyzes the current composition, and predicts text guidance describing how to improve it. (2) *Generation*. The predicted text guidance is first appended to the inputs. Starting from pure noise tokens, the model progressively denoises to generate latent tokens. After VAE decoding, a well-composed example image is obtained.

5. Experiments

5.1. Metrics and Details

Evaluation of example images. Assessing composition is non-trivial. Although we have trained a composition assessment model, it was used in our dataset construction and thus would bias the evaluation. As noted in [23, 68], humans find it easier to compare two images than to rate a single one. Therefore, we evaluate each generated example by comparing it with both the original and ground-truth images. The comparison is conducted by GPT-5 [43] and humans, and the two win rates are reported as the metric. We manually select and carefully examine 200 to 300 samples for each task to construct a benchmark to compute the metric.

Evaluation of text guidance. We evaluate the consistency between the text guidance and the corresponding example images. Specifically, we input the original image, the model-improved example image, and the model-predicted text guidance into GPT-5 [43], and request an evaluation score indicating how accurately the text guidance describes the change from the original image to the example image.

Implementation details. We follow the training setup of Bagel [9] to train our model. The VAE encoder and decoder are kept frozen, while the ViT encoder and the main model remain trainable. Training is performed on 8 NVIDIA A100 GPUs using the AdamW optimizer [21] with a batch size of 8, a learning rate of 2e-5, and 50K training steps. An exponential moving average with a decay rate of 0.9999 is applied to stabilize training. The two loss terms, $\mathcal{L}_{text}$ and $\mathcal{L}_{img}$, are assigned equal weights. All images are resized to a 512 shorter side while preserving the aspect ratio. During inference, the number of image generation steps is set to 30.

5.2. Comparison Results

Baselines. We compare with state-of-the-art editing models, including Kontext [3] and Qwen-Image-Edit [60]. The

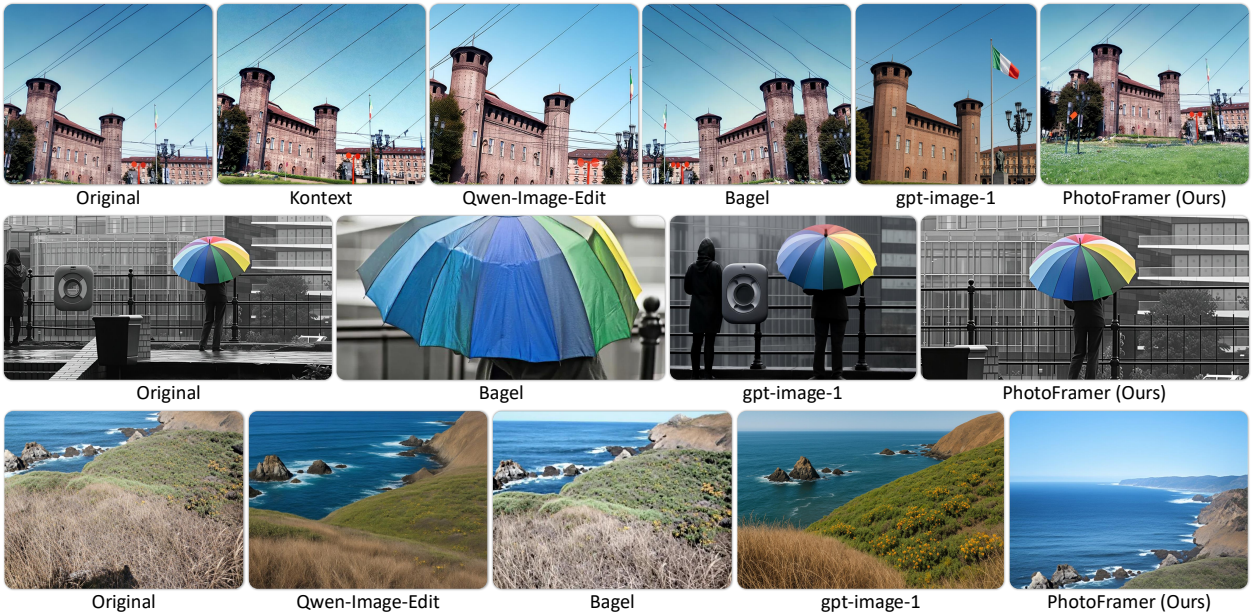

Figure 7. Qualitative comparison between our PhotoFramer and baseline methods. Open-source editing models fail to improve composition. The gpt-image-1 [42] could generate well-composed images but with low fidelity (*i.e.*, it alters semantic details of the original image).

Table 3. Quantitative results of model-generated example images on our benchmark. We compare the model-generated example images with their original images / ground-truth images, evaluated by both GPT-5 [43] and humans. The win rate (%) is reported as the metric. We also report image quality assessment (with DeQA-Score [69]) and image aesthetic assessment (with Q-Align [61]) results for reference.

Method	Shift Task		Zoom-in Task[1]		View-change Task		Quality	Aesthetic
	GPT-5 [43]	Human	GPT-5 [43]	Human	GPT-5 [43]	Human	DeQA [61]	Q-Align [61]
Kontext [3]	39.88 / 12.27	49.69 / 4.94	- / 15.52	- / 5.17	46.74 / 15.76	48.37 / 5.98	3.88	3.13
Qwen-Image-Edit [60]	46.01 / 16.56	48.43 / 10.49	- / 39.65	- / 13.79	70.65 / 36.96	61.96 / 20.65	4.03	**3.29**
Bagel [9]	27.61 / 14.73	38.36 / 8.02	- / 22.41	- / 12.07	47.28 / 14.13	64.13 / 15.22	3.87	3.08
gpt-image-1 [42]	69.93 / 33.99	68.46 / 22.37	- / 52.73	- / 27.27	**84.61 / 51.65**	81.52 / 41.30	3.97	3.26
PhotoFramer (Ours)	**80.37 / 35.58**	**88.05 / 43.83**	- / **67.24**	- / **48.28**	82.07 / 50.54	**85.87 / 47.28**	**4.07**	3.17

[1] For the zoom-in task, we **exclude** Zoomed *vs*. Original because the pair is trivially detectable (*i.e.*, through scale/cropping cues), which induces type preference and inflates win rates. We therefore report only Zoomed *vs*. Ground-Truth win rate.

original Bagel [9] is also included, along with the powerfully proprietary gpt-image-1 [42]. We adopt Bagel's reasoning-based editing mode, which first generates textual edit instructions and then edits the image accordingly.

Quantitative results of example images are presented in Tab. 3. First, open-source editing models fail to improve composition, exhibiting low win rates against both the original and ground-truth images. Second, in the view-change task, where the models have greater freedom in generation, their performance is substantially improved. Third, gpt-image-1 achieves substantially better results than open-source models, demonstrating strong generalization ability to this new task. However, gpt-image-1 often alters semantic details of the original image, as depicted in Fig. 7. Finally, our PhotoFramer outperforms open-source methods with the highest win rates, and matches or even surpasses gpt-image-1. Tab. 3 shows that our model not only improves composition but also preserves high-level image quality and aesthetics. The qualitative examples in Fig. 1 and Fig. 7 show that PhotoFramer effectively enhances composition while maintaining high fidelity to the original content.

Table 4. Quantitative results of consistency between the model-predicted text guidance and corresponding example images.

Task	Shift	Zoom-in	View-change	Average
Bagel [9]	77.01	84.82	87.47	83.10
PhotoFramer (Ours)	**91.96**	**92.59**	**91.52**	**92.02**

Evaluation results of text guidance are provided in Tab. 4. The original Bagel model has exhibited reasonable consistency between text guidance and the corresponding example images across the three composition instruction sub-tasks. After finetuning on our constructed datasets, this consistency is further and stably improved (92.02% *vs*. 83.10%).

5.3. Ablation Studies and Discussions

Text guidance alone is not enough. A natural question arises for the unified framework: can we rely solely on the text guidance and feed it into other editing models to generate example images? As shown in Fig. 8, our generated text guidance cannot be directly utilized by Qwen-Image-Edit, though it employs a Large Language Model (LLM) as

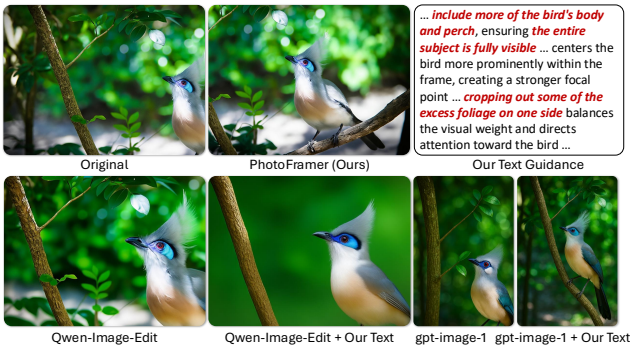

Figure 8. Our generated text guidance cannot be directly utilized by Qwen-Image-Edit, even though it employs an LLM as the text encoder. In contrast, gpt-image-1 benefits from text guidance (*i.e.*, including the entire body of the bird), albeit with lower fidelity.

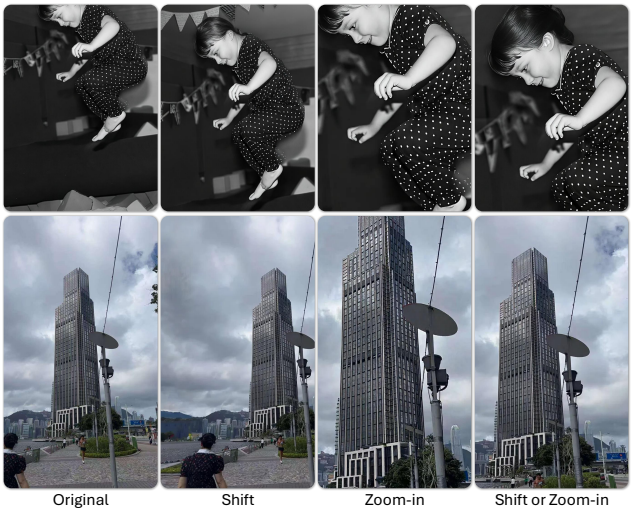

Figure 9. Illustration of auto prompt. Given an auto prompt "refine the composition using shift or zoom-in", our model does not merely select one task type but adaptively fuses multiple operations to produce a better composition. Text guidance is omitted.

the text encoder. Moreover, directly using our text guidance even degrades the fidelity of example images. In contrast, when provided with our text guidance (*i.e.*, "include more of the bird's body and perch"), gpt-image-1 successfully includes the entire bird, showing strong instruction-following ability. However, its fidelity remains unsatisfactory.

Example image alone is not enough. First, as depicted in Fig. 10, textual guidance plays a crucial role in generating example images. Even revising a few key words could lead to dramatically different results. Second, we train Bagel using only image pairs without text guidance. As illustrated in Fig. 11, without textual input, the model fails to remove the foreground fence, although it successfully includes the sky. In contrast, when trained with text guidance, the model explicitly learns to "remove the fence" and successfully follows this instruction to generate a better example image. Third, we fine-tune Kontext (which does not support textual training) using our collected image pairs. As shown in

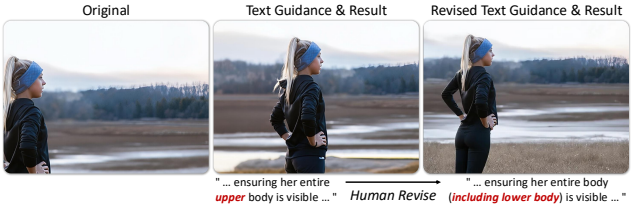

Figure 10. Text guidance is important for image generation. If we manually revise a few key words (*i.e.*, remove "upper", add "including lower body"), the generated image will be quite different.

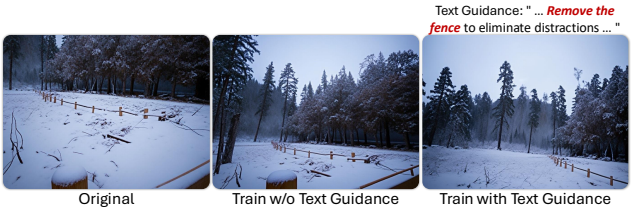

Figure 11. Training without text guidance fails to remove the fence distractions, although it successfully includes the sky, whereas training with text guidance predicts to "remove the fence" in textual output and generates a well-composed image without fence.

Figure 12. Finetuned Bagel (*i.e.*, on both textual and visual data) successfully includes the *whole* wooden house as in the text guidance, outperforming finetuned Kontext (*i.e.*, on visual data only).

Fig. 12, Kontext can partially shift the wooden house toward the center but fails to include it entirely. In contrast, the finetuned Bagel predicts to "include the *whole* wooden structure" and generates a well-composed image accordingly.

Task prompt. As illustrated at the top of Fig. 9, different task prompts lead PhotoFramer to apply different operations to improve composition. Notably, as shown at the bottom of Fig. 9, when specific tasks fail, the auto task can still produce strong results. The original image was randomly taken in a city and contains a distraction (*i.e.*, the back of a half-visible woman). The shift task mistakenly treats this distraction as the main subject and attempts to center it, while the zoom-in task crops too tightly, cutting off the top of the building. With the auto prompt, where the model can adaptively apply shift and/or zoom-in, PhotoFramer effectively fuses multiple adjustments to achieve a better composition.

6. Conclusions

We introduce PhotoFramer, a multi-modal composition instruction model built upon a hierarchical task paradigm, curated datasets, and a unified understanding–generation framework, to guide photographic composition through actionable textual instructions and example image generation.

Acknowledgment. This research work was supported by National Natural Science Foundation of China (Grant No. 62276251), RGC Early Career Scheme (ECS) No. 24209224, and the Joint Lab of CAS-HK.

References

[1] Hadi AlZayer, Hubert Lin, and Kavita Bala. AutoPhoto: Aesthetic photo capture using reinforcement learning. In *International Conference on Intelligent Robots and Systems (IROS)*, 2021. 4, 5, 17

[2] 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*, 2025. 2, 3, 4, 5, 13, 14

[3] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 6, 7

[4] Mingdeng Cao, Xuaner Zhang, Yinqiang Zheng, and Zhihao Xia. Instruction-based image manipulation by watching how things move. In *CVPR*, 2025. 2

[5] 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. *arXiv preprint arXiv:2507.14533*, 2025. 2

[6] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *WACV*, 2017. 2, 3

[7] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *ACM MM*, 2017. 2, 4, 5

[8] Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Mobile computational photography: A tour. *Annual Review of Vision Science*, 2021. 1

[9] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2, 3, 6, 7

[10] Unsplash Developers. Unsplash lite dataset 1.3.0, 2020. 2, 5, 16

[11] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *ACM MM*, 2014. 2, 3, 15

[12] Farshid Farhat, Mohammad Mahdi Kamani, Sahil Mishra, and James Z Wang. Intelligent portrait composition assistance: Integrating deep-learned models and photography idea retrieval. In *ACM MM Workshops*, 2017. 2

[13] Farshid Farhat, Mohammad Mahdi Kamani, and James Z Wang. CAPTAIN: Comprehensive composition assistance for photo taking. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2022. 2

[14] Michael Freeman. *The Photographer's Eye Digitally Remastered 10th Anniversary Edition: Composition and Design for Better Digital Photos*. Routledge, 2017. 2

[15] Michael Freeman. *The Photographer's Eye Digitally Remastered 10th Anniversary Edition: Composition and Design for Better Digital Photos*. Routledge, 2017. 3

[16] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 2

[17] Guanjun Guo, Hanzi Wang, Chunhua Shen, Yan Yan, and Hong-Yuan Mark Liao. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *IEEE TMM*, 2018. 2

[18] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *CVPR*, 2021. 2

[19] James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning subject-aware cropping by outpainting professional photos. In *AAAI*, 2024. 2

[20] Andrey Ignatov, Georgii Perevozchikov, Radu Timofte, Cheng Li, Lian Liu, Jun Cao, Heng Sun, Wu Pan, Song Wang, KeQiang Yu, et al. Learned smartphone ISP on mobile GPUs, mobile AI 2025 challenge: Report. In *CVPR Workshops*, 2025. 1

[21] Loshchilov Ilya and Hutter Frank. Decoupled weight decay regularization. In *ICLR*, 2019. 6, 14

[22] Gengyun Jia, Huaibo Huang, Chaoyou Fu, and Ran He. Rethinking image cropping: Exploring diverse compositions from global views. In *CVPR*, 2022. 2

[23] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *ECCV*, 2020. 6

[24] Diederik P Kingma. Adam: A method for stochastic optimization. In *ICLR*, 2015. 13

[25] Black Forest Labs. FLUX, 2024. 6

[26] Michael Langford. *Basic photography*. Routledge, 2013. 2

[27] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Photographic composition classification and dominant geometric element detection for outdoor scenes. *Journal of Visual Communication and Image Representation*, 2018. 2, 5, 13

[28] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-RL: Aesthetics aware reinforcement learning for image cropping. In *CVPR*, 2018. 2

[29] Debang Li, Junge Zhang, and Kaiqi Huang. Learning to learn cropping models for different aspect ratio requirements. In *CVPR*, 2020. 2

[30] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *CVPR*, 2020. 2

[31] Jiawan Li, Fei Zhou, Zhipeng Zhong, Jiongzhi Lin, and Guoping Qiu. Towards smart point-and-shoot photography. In *CVPR*, 2025. 2

[32] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-Insight: Understanding image quality via visual reinforcement learning. In *NeurIPS*, 2025. 5, 13

[33] Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM TOG*, 2025. 4, 15

[34] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024. 2, 5, 14

[35] Boyang Liu, Yifan Hu, Senjie Jin, Shihan Dou, Gonglei Shi, Jie Shao, Tao Gui, and Xuanjing Huang. Unlocking the essence of beauty: Advanced aesthetic reasoning with relative-absolute policy optimization. *arXiv preprint arXiv:2509.21871*, 2025. 2

[36] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 1989. 13

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 2

[38] Xiaoyu Liu, Ming Liu, Junyi Li, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Beyond image borders: Learning feature extrapolation for unbounded image composition. In *ICCV*, 2023. 2

[39] Chamin Morikawa, Michihiro Kobayashi, Masaki Satoh, Yasuhiro Kuroda, Teppei Inomata, Hitoshi Matsuo, Takeshi Miura, and Masaki Hilaga. Image and video processing on mobile devices: a survey. *The Visual Computer*, 2021. 1

[40] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 4, 5, 13

[41] OpenAI. GPT-4V(ision) system card, 2023. 2

[42] OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[43] OpenAI. GPT-5 system card, 2025. 6, 7

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 4

[45] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U^2-Net: Going deeper with nested u-structure for salient object detection. *PR*, 2020. 4

[46] 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 *ICML*, 2021. 4

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[48] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepseekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 5, 13

[49] Yukun Su, Yiwen Cao, Jingliang Deng, Fengyun Rao, and Qingyao Wu. Spatial-semantic collaborative cropping for user generated content. In *AAAI*, 2024. 2

[50] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *ICLR*, 2024. 2

[51] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[52] DXOMARK Team. DXOMARK - quality testing, scores and reviews, 2025. 1

[53] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. MetaMorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024. 2, 3

[54] 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. 6

[55] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *AAAI*, 2020. 2

[56] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025. 5

[57] Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *ICCV*, 2017. 2

[58] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *CVPR*, 2018. 2, 3, 4, 5, 12, 15

[59] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, 2025. 2

[60] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 6, 7

[61] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *ICML*, 2024. 3, 4, 5, 7

[62] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal llm. In *ICML*, 2024. 2

[63] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. In *NeurIPS*, 2025. 5, 13

[64] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025. 2

[65] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *CVPR*, 2013. 2, 3, 15

[66] Guo-Ye Yang, Wen-Yang Zhou, Yun Cai, Song-Hai Zhang, and Fang-Lue Zhang. Focusing on your subject: Deep subject-aware image composition recommendation networks. *Computational Visual Media*, 2023. 2, 3, 15

[67] Zhiyuan You, Jinjin Gu, Zheyuan Li, Xin Cai, Kaiwen Zhu, Chao Dong, and Tianfan Xue. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*, 2024. 3

[68] 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 *ECCV*, 2024. 3, 6

[69] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *CVPR*, 2025. 5, 7

[70] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014. 15

[71] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *CVPR*, 2019. 2

[72] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE TPAMI*, 2020. 2, 3, 4, 5, 13, 15

[73] Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. In *BMVC*, 2021. 2, 4, 5, 13

[74] Bo Zhang, Li Niu, Xing Zhao, and Liqing Zhang. Human-centric image cropping with partition-aware and content-preserving features. In *ECCV*, 2022. 2

[75] Ke Zhang, Tianyu Ding, Jiachen Jiang, Tianyi Chen, Ilya Zharkov, Vishal M Patel, and Luming Liang. ProCrop: Learning aesthetic image cropping from professional compositions. *arXiv preprint arXiv:2505.22490*, 2025. 2

[76] Xiaoyan Zhang, Zhuopeng Li, Martin Constable, Kap Luk Chan, Zhenhua Tang, and Gaoyang Tang. Pose-based composition improvement for portrait photographs. *IEEE TCSVT*, 2018. 2

[77] Zhaoran Zhao, Peng Lu, Anran Zhang, Peipei Li, Xia Li, Xuannan Liu, Yang Hu, Shiyi Chen, Liwei Wang, and Wenhao Guo. Can machines understand composition? dataset and benchmark for photographic image composition embedding and understanding. In *CVPR*, 2025. 2

[78] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024. 2

[79] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2