

Appendix

A. More Results

We have added more qualitative results in Figs. A11 to A13, including the poorly composed input, the predicted text guidance, and the model-generated well-composed image, for all three sub-tasks. In the predicted text guidance, actionable suggestions are highlighted in **blue**, while hallucinated descriptions are marked in **red**. Overall, our PhotoFramer could provide clear and practically useful composition suggestions that correspond well to the generated example well-composed images. However, in some cases, small hallucinations still occur. For instance, in the shift task, the model may confuse left and right directions, leading to incorrect spatial descriptions.

We also test if one iteration of view improvement would be sufficient to get the desired result or whether iterative approach is needed. In many cases, a single round of guidance is enough to yield a clear improvement. Furthermore, as illustrated in Fig. A1, our approach can be applied to refine the image composition iteratively.



Figure A1. Our PhotoFramer can be applied to refine the image composition iteratively.

B. More Details of Dataset Construction

In this section, we provide more details of our dataset construction, including the optimization to obtain composition scores for CPC dataset (Sec. B.1), composition assessment model training (Sec. B.2), details of view-change pairs collection (Sec. B.3), more dataset statistics (Sec. B.4), and task prompt details (Sec. B.5).

B.1. Score Optimization of CPC Dataset

As stated in Sec. 3.2.1 of the main paper, the composition scores are required for each crop to construct shift or zoom-in pairs. However, the CPC dataset [58] contains only selected good and best crops, which cannot be directly used to construct the pairs. Therefore, we build a mathematical model to infer scores for all crops.

Annotation of CPC dataset. Given an original image and its N crops ($N=24$ in CPC), one annotator first selects 8~20 good crops, and then chooses the 3 best crops from these good ones. Each image is annotated by 6 annotators.

By averaging the annotations from the 6 annotators, we obtain the probability that crop i is good, p_i^{good} , and the probability that it belongs to the top-3 set, p_i^{top3} .

Assumption. Annotators are required to select 8~20 good crops out of all 24 crops, where good crops constitute a relatively large proportion. Therefore, we assume that a crop i is considered good if its score s_i exceeds a threshold, and this decision *depends only on its own quality without pairwise comparison*. In contrast, selecting the top-3 crops requires careful comparison among candidates. Thus, the decision necessarily *depends on relative scores between crops*.

Mathematical modeling. According to the above assumption, whether a crop is considered good or not *depends only on its own quality without pairwise comparison*. Therefore, we directly model the probability of a crop being good as a sigmoid function of its own score:

$$\hat{p}_i^{\text{good}} = \sigma(s_i), \quad (\text{A1})$$

where $\sigma(\cdot)$ denotes the sigmoid activation function.

For the top-3 crops, careful comparison among crop candidates must be required. Therefore, we first model the probability of each crop being selected as the top-1 using a softmax (*i.e.*, comparison) over all crop scores:

$$\hat{p}^{\text{top1}} = \text{softmax}(\alpha[s_0, \dots, s_i, \dots, s_N]) \in \mathbb{R}^N, \quad (\text{A2})$$

where α is a scaling factor, and it is set to 2. We can view this problem as a sampling problem, where p_i^{top1} denotes the probability of sampling crop i . Selecting the best top-3 can then be formulated as sampling without replacement. The probability that crop i is selected into the top-3 is:

$$\begin{aligned} \hat{p}_i^{\text{top3}} &= p_i^{\text{top1}} \\ &\quad (\rightarrow \text{ crop } i \text{ is chosen first}) \\ &+ p_i^{\text{top1}} \sum_{j \neq i} \frac{p_j^{\text{top1}}}{1 - p_j^{\text{top1}}} \\ &\quad (\rightarrow \text{ crop } i \text{ is chosen second}) \quad (\text{A3}) \\ &+ p_i^{\text{top1}} \sum_{j \neq i} \sum_{\substack{k \neq i \\ k \neq j}} \frac{p_j^{\text{top1}} p_k^{\text{top1}}}{(1 - p_j^{\text{top1}})(1 - p_j^{\text{top1}} - p_k^{\text{top1}})} \\ &\quad (\rightarrow \text{ crop } i \text{ is chosen third}) \end{aligned}$$

This formulation is difficult to compute and optimize. Therefore, considering that there are a total of 24 crops for each image and $24 \gg 3$, we approximate the process of “sampling the top-3 without replacement” as “3 independent samplings with replacement”.

$$\hat{p}_i^{\text{top3}} = 1 - (1 - p_i^{\text{top1}})^3. \quad (\text{A4})$$

Mathematically, Eq. (A3) can be approximated by Eq. (A4) when N is large and $\max(p_i^{\text{top1}})$ is relatively small. First,

in CPC, each image contains 24 crops, and $N=24$ is sufficiently larger than the sampling size of 3. Second, based on the human-annotated $p_i^{\text{top}3}$, we observe that in most cases no single crop is selected among the top-3 by more than 4 out of 6 annotators, indicating that multiple crops are competitive. Thus, no crop clearly dominates others, *i.e.*, $\max(p_i^{\text{top}1})$ remains small, supporting that Eq. (A3) can be well approximated by Eq. (A4).

Loss function and optimization. We treat the composition scores s_i as learnable parameters, from which we estimate the probability that crop i is good, \hat{p}_i^{good} , and the probability that it belongs to the top-3 set, $\hat{p}_i^{\text{top}3}$. We define the loss as the discrepancy between the estimated probabilities and human-annotated ground truth:

$$\mathcal{L} = \sum_{i=1}^N \left((\hat{p}_i^{\text{good}} - p_i^{\text{good}})^2 + \beta (\hat{p}_i^{\text{top}3} - p_i^{\text{top}3})^2 \right), \quad (\text{A5})$$

where β is a weighting factor, and it is set to 2 by default to emphasize the top-3 term.

We optimize the scores for each original image in the CPC dataset independently. We first apply the L-BFGS [36] optimizer (lr=1.0, max_epochs=10, max_iter=200, history_size=10), which is efficient (less than 1 second per image). However, it can be unstable, leading to optimization failures. Therefore, when the final loss $\mathcal{L} > 0.5$, we switch to the Adam [24] optimizer (lr=2e-3, max_epochs=5000), which is slower (20~30 seconds per image) but more stable. Overall, the average loss converges to around 0.36. The mean absolute errors are 0.0602 between \hat{p}_i^{good} and p_i^{good} , and 0.0426 between $\hat{p}_i^{\text{top}3}$ and $p_i^{\text{top}3}$, indicating sufficiently accurate probability estimation for practical use.

Post processing. The optimized composition scores are clipped and normalized to the range [1,5] for further use.

B.2. Details of Composition Assessment Model

Composition assessment dataset collection. To train the composition assessment model, we primarily use composition scoring datasets, CADB [73] and GAIC [72], where each image is annotated with a composition score. We additionally incorporate composition classification datasets, CADB [73] and KU-PCP [27], and the aesthetic assessment dataset AVA [40], due to their strong relevance to the composition evaluation. The composition scores or aesthetic scores of each dataset are normalized to the range of [1, 5]. There are 13 composition categories in CADB dataset, *i.e.*, [center, curved, diagonal, fill the frame, golden ratio, horizontal, pattern, radial, rule of thirds, symmetric, triangle, vanishing point, vertical]. KU-PCP dataset contains 9 composition classes including [center, curved, diagonal, horizontal, pattern, rule of thirds, symmetric, triangle, vertical].

We perform data resampling strategy on the AVA and GAIC datasets to mitigate data imbalance and redundancy.

Table A1. Re-sampling AVA dataset to increase the proportion of low (*i.e.*, [1,2]) and high (*i.e.*, [4,5]) score range.

Score range	[1,2)	[2,3)	[3,4)	[4,5]
Original	2,051 / 0.87%	91,401 / 38.80%	137,946 / 58.55%	4,200 / 1.78%
Re-sampled	2,051 / 4.75%	17,408 / 40.28%	19,562 / 45.26%	4,200 / 9.72%

Table A2. Statistics of our collected and re-sampled datasets to train the composition assessment model.

	CADB	Assessment		Classification	
		GAIC	AVA	CADB	KU-PCP
Train	8,547	48,140	43,221	8,547	4,244
Test	950	8,472	4,557	950	-

First, the AVA dataset contains an excessive number of mid-quality images (*i.e.*, 97.35% in the [2,4) score range), while low-quality (*i.e.*, [1,2)) and high-quality (*i.e.*, [4,5]) samples account for only 0.87% and 1.78%, respectively, which may bias model training. Therefore, we re-sample AVA to increase the proportions of low- and high-quality images, as illustrated in Tab. A1. Second, the GAIC dataset includes 288K cropped images with MOS labels, but these are derived from only 3,336 original images, each contributing over 80 crops, resulting in insufficient data diversity. To address this, we sample 20% of GAIC, ensuring a diverse MOS distribution, which yields approximately 56K samples for training and evaluation. The final dataset statistics are summarized in Tab. A2.

Composition assessment model training using GRPO. Following Q-Inspire [32] and VisualQuality-R1 [63], we adopt Qwen2.5-VL-7B [2] as the base model and train it using the GRPO [48] reinforcement learning algorithm. Specifically, for each question, we request the model to first output the thinking process in `<think></think>` tags and then output the final answer in `<answer></answer>` tags. We repeatedly request the model by N times to obtain N outputs, $\{o_1, o_2, \dots, o_N\}$. Then, for each output, we extract the answer between `<answer></answer>` tags, and calculate rewards $\{r_1, r_2, \dots, r_N\}$ by comparing the answers and ground truth. By calculating the mean and standard deviation of the rewards, the relative advantages of each response can be obtained as follows.

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_N\})}{\text{std}(\{r_1, r_2, \dots, r_N\})}, \quad (\text{A6})$$

where \hat{A}_i denotes the normalized relative advantage (quality) of the i -th response. Overall, GRPO guides the policy model to prioritize higher-quality responses that receive higher reward values within each group. After obtaining \hat{A}_i , GRPO computes the ratio between the probabilities of the same response under the updated policy $\pi_{\theta_{\text{new}}}$ and the previous policy $\pi_{\theta_{\text{old}}}$, denoted as ρ_i . To prevent excessively large policy updates and stabilize training, ρ_i is constrained

within the range $[1 - \delta, 1 + \delta]$. In addition, to maintain proximity to the reference distribution π_{ref} , a KL-divergence penalty weighted by β is introduced. Finally, the optimization objective of GRPO can be formulated as follows:

$$\mathcal{J}(\theta) = \mathbb{E}_{q \sim \mathcal{Q}, o_i \sim \pi_{\theta_{\text{old}}}} \left\{ \left[\min \left(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1 - \delta, 1 + \delta) \hat{A}_i \right) \right] - \beta \mathbb{D}_{\text{KL}}(\pi_{\theta_{\text{new}}} \parallel \pi_{\text{ref}}) \right\}, \quad (\text{A7})$$

where $\rho_i = \pi_{\theta_{\text{new}}}(o_i | q) / \pi_{\theta_{\text{old}}}(o_i | q)$, \mathcal{Q} denotes the set of candidate questions, and \mathbb{D}_{KL} is the KL regularization term. The reference policy π_{ref} is typically a frozen pre-trained MLLM. Overall, GRPO balances consistent policy updates with strong reward signals, enabling stable yet effective optimization.

Reward calculation. To optimize Eq. (A7), we need to calculate reward r_i for each output o_i .

- *Format reward* evaluates whether the reasoning steps are properly enclosed within the `<think></think>` tags, and whether the final answer is correctly enclosed within the `<answer></answer>` tags. The format reward r_i^{format} is set to 1 if the i -th response satisfies above conditions; otherwise, it is set to 0.

- *Score reward* evaluates whether the predicted composition or aesthetic score is accurate. We first extract the model-predicted score within the `<answer></answer>` tags, then compute its error with respect to the ground-truth score. If the error is equal to or smaller than a threshold σ , the score reward r_i^{score} for the i -th response is set to 1; otherwise, it is set to 0. The threshold σ is set as 0.4 empirically.

- *Classification reward* evaluates whether the predicted composition categories match the ground-truth categories. We provide the model with all possible composition categories and ask it to select one to three of the most suitable ones (since the ground-truth annotations may contain one to three types). Let the ground truth contain n composition types, the model predict m types, and let k be the number of correctly predicted types. The classification reward is then computed as $r_i^{\text{class}} = k / \max(m, n)$.

Finally, the overall reward for the composition or aesthetic assessment task is defined as:

$$r_i = \begin{cases} r_i^{\text{format}} + r_i^{\text{score}}, & \text{if } r_i^{\text{format}} = 1, \\ 0, & \text{otherwise,} \end{cases} \quad (\text{A8})$$

and the overall reward for composition classification task is:

$$r_i = \begin{cases} r_i^{\text{format}} + r_i^{\text{class}}, & \text{if } r_i^{\text{format}} = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (\text{A9})$$

With the final reward r_i for each output o_i , we then optimize Eq. (A7) to train the composition assessment model.

Implementation details. We adopt Qwen2.5-VL-7B [2] as the base model, with all model components, including the vision encoder, vision–text connector, and large language model, kept trainable. To ensure the data balance, the CADB and KU-PCP datasets are duplicated by 5 times. For each question, the number of generated responses N is set to 8. Training is performed for 1 epoch with a batch size of 32 on 8 NVIDIA A6000 GPUs using 4 gradient-accumulation steps. We employ the AdamW optimizer [21] with a learning rate of $1e-6$. The coefficients β and ϵ in Eq. (A7) are set to 0.04 and 0.2, respectively.

Composition assessment results. Quantitative results of our composition assessment model have been reported in Tab. 2 of the main paper. Here, we provide additional qualitative examples in Fig. A9 and Fig. A10. The results show that our model accurately understands compositional structure and produces reliable composition scores accompanied by detailed and coherent reasoning.

B.3. Details of View-change Pairs Collection

View-change pairs sampled from the multi-view 3D dataset. We visualize the sampled View-change pairs from the DL3DV [34] dataset in Fig. A2. As shown, even the good-composition images in DL3DV reflect ordinary, casually captured viewpoints rather than expert-level compositions. This motivates the introduction of the degradation model, which is then applied to expert-taken photos to further construct higher-quality multi-view pairs.

Text–vision joint training of the degradation model.

As illustrated in Fig. 11 of the main paper, incorporating text guidance leads to higher-quality results compared with image-only training. Therefore, we adopt the same text–vision joint training strategy for the degradation model. Let the good image be denoted as `I_good`, the corresponding poor image as `I_poor`, the task prompt as `T_task`, the predicted text guidance as `T_guide`, and the degradation model as `f()`. The degradation model is formulated as: `I_poor, T_guide = f(I_good, T_task)`.

To train such a degradation model, we need to construct `<T_task, I_poor, I_good, T_guide>` samples. The image pairs `<I_poor, I_good>` are collected from multi-view datasets, as illustrated in Fig. 5 of the main paper. The task prompt `T_task`, analogous to the three sub-tasks, is randomly sampled from a set of predefined template sentences such as “Change the viewpoint to *worsen* the composition”. Finally, we input each image pair into the vision language model Qwen2.5-VL-32B [2] to annotate the text guidance `T_guide`, which describes how the good image is transformed into the poor one.

With the above collected data, we could finally train the degradation model using the same text–vision joint training strategy described in Sec. 4 of the main paper. Some qualitative degradation examples generated by our degradation



Figure A2. View-change pairs sampled by our composition assessment model from the DL3DV dataset.

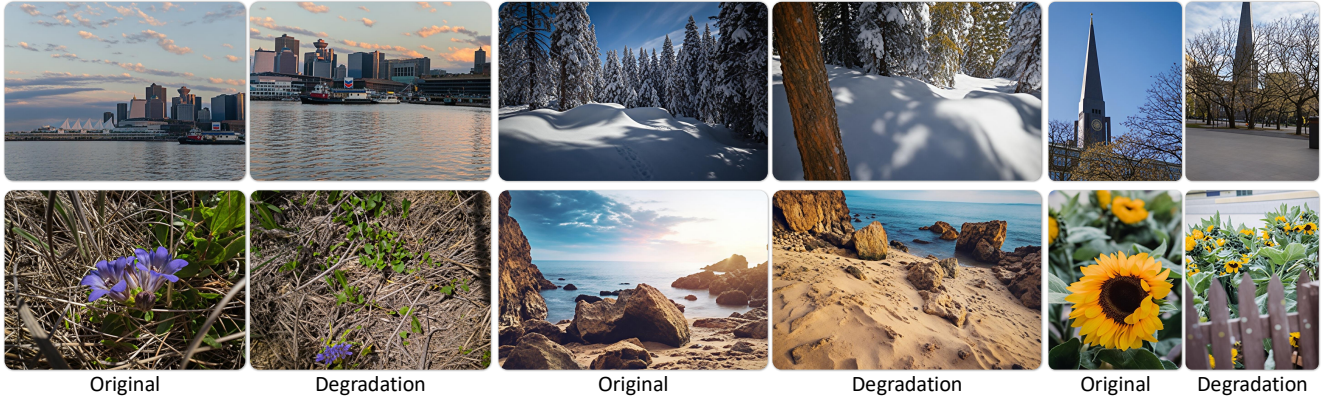


Figure A3. Degradation examples generated by our degradation model.

model are provided in Fig. A3.

Image restoration to enhance degraded outputs. As stated in the main paper, we apply the degradation model to human-taken photos produces poor-composition images. However, these degraded outputs may occasionally suffer from low visual quality with noticeable artifacts, as shown in Fig. A4. To mitigate this issue, we employ the state-of-the-art image restoration model HYPIR [33] to enhance the visual quality of degraded images. Specifically, we first down-sample each image by a factor of 4 and then use HYPIR to up-sample it by the same factor, which further suppresses generative artifacts. As illustrated in Fig. A4, HYPIR substantially improves image fidelity, yielding cleaner and more natural-looking degraded images.

B.4. Dataset Statistics

Image statistics. We have provided a rough statistic of our collected dataset in the Tab. 1 of the main paper. Here we perform a more detailed statistic of each sub-task in Tabs. A3 to A5. Since the shift and zoom-in pairs are sampled from the original images in existing cropping datasets including GAIC [72], CPC [58], SACD [66], FLMS [11], FlickrCrop [70], CUHKCrop [65], both the original images

Table A3. Statistics of training / test shift dataset.

	GAIC	CPC	Other
# Original	2,613 / 277	6,128 / 682	621 / -
# Pairs	72,710 / 4,578	80,486 / 6,432	698 / -

Table A4. Statistics of training / test zoom-in dataset.

	GAIC	CPC	SACD	FLMS	FlickrCrop	CUHKCrop
# Original	1,603 / 147	4,099 / 459	796 / 70	338 / -	110 / -	43 / -
# Pairs	3,830 / 327	7,426 / 869	813 / 70	686 / -	110 / -	51 / -

Table A5. Statistics of training / test view-change dataset.

	DL3DV	Unsplash Lite	Our Collected
# Pairs	13,335 / 710	7,289 / 1,000	4,340 / 719

and pairs are included in the statistics. In Tab. A3, besides the collected shift pairs from GAIC and CPC, we also collect a small number of shift pairs from FlickrCrop and other resources (*e.g.*, AIGC creation), denoted as “Other”.

Text statistics. We perform statistics on the text guidance length of three sub-tasks in Tab. A6. The average word length is about 100 words, providing sufficiently rich yet concise guidance for composition refinement. The word

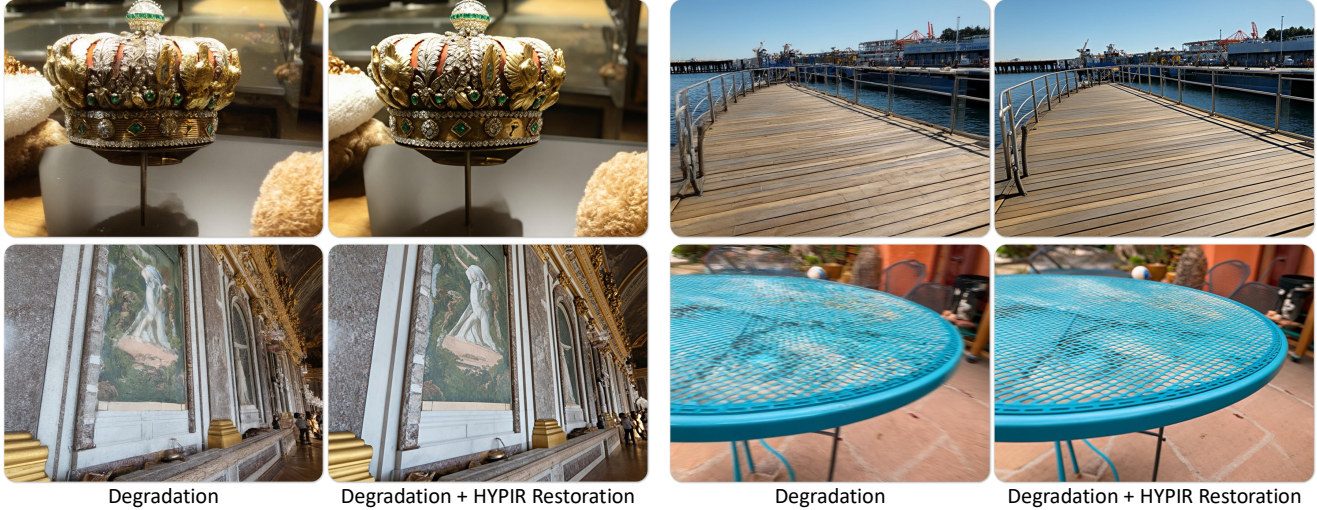


Figure A4. We use a state-of-the-art restoration model HYPIR to enhance the quality of degraded images.

Table A6. Statistics of the text guidance length.

	Shift	Zoom-in	View-change
# Word Length	94.65	102.60	97.06
# String Length	613.43	663.16	632.77

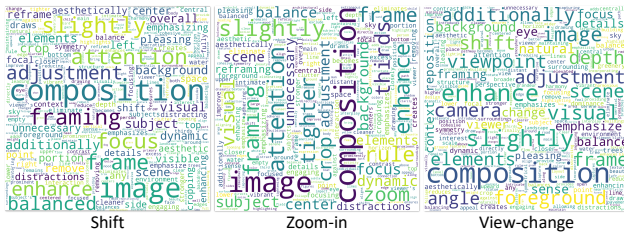


Figure A5. Word cloud of the text guidance.

cloud of the text guidance is depicted in Fig. A5, where high-frequency words such as “composition”, “balance”, “framing”, “angle”, “viewpoint”, “tighter”, and “depth” are all closely related to compositional principles.

B.5. Other Details

Task prompt. The task prompts for all three sub-tasks are provided in Tab. A7. Moreover, the prompts used for the two auto tasks are listed in Tab. A8.

C. Limitations and Discussions

First, the degradation model may fail to worsen the good composition. This issue is particularly common for expert-taken photos in the Unsplash Lite [10] dataset, where the degraded outputs can remain well-composed, as illustrated in Fig. A6. Fortunately, we observe empirically that including such $\langle \text{good}, \text{good} \rangle$ pairs does not harm the model performance. Moreover, since the users may also input such already well-composed images, we intentionally retain these samples rather than filtering them out.



Figure A6. The degradation model may fail to produce a poorly composed image when given an expert-taken photo, in which case the degraded output can still appear well-composed.

Second, the degradation model may change the image excessively, producing pairs that are no longer semantically consistent. As shown in Fig. A7, the degraded output deviates notably from the original content. Although we perform data filtering, a small number of such inconsistent pairs may still remain. Future work should explore a more controllable degradation model to reduce inconsistency.

Third, the generated images may exhibit relatively low visual quality, as depicted in Fig. A4. However, image quality is not the primary focus of our work, and we are concerned chiefly with composition. Improving the quality of generated images can be explored in our future work.

Fourth, we emphasize that synthetic data provides only an initial foundation for training composition instruction models. While it enables large-scale supervision and controlled variation, it cannot fully replace real-world images

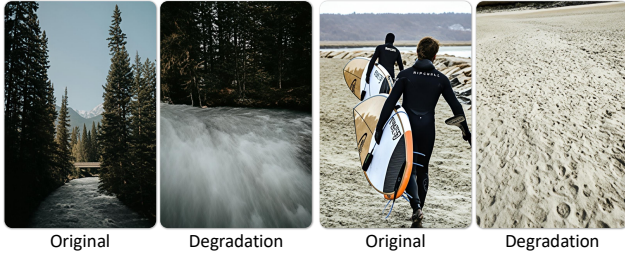


Figure A7. The degradation model may alter the image excessively, leading to semantically inconsistent pairs.



Figure A8. A failure case where our PhotoFramer misinterprets a painting as a real photograph and attempts to refine the composition by removing the “background” (*i.e.*, the painting frame).

captured by everyday mobile users, which exhibit diverse shooting habits, device characteristics, and natural composition errors. Constructing such a real, human-captured dataset will be crucial for advancing composition instruction and remains an important direction for future work. We present a failure case in Fig. A8 where PhotoFramer misinterprets a painting as a real photograph and attempts to refine the composition by removing the perceived “background” (*i.e.*, the painting frame). This demonstrates the limitations of our small-scale datasets.

Finally, although it is best to deploy the model on a robot (like [1]) to automatically capture photos, real-world robot deployment remains challenging due to the compute and latency constraints of running a 14B model on-device. Building a reliable hardware-in-the-loop platform is beyond the scope of this paper and is left for future work.

Table A7. **Task prompt** of three sub-tasks.

#	Shift Prompt
1	Shift the framing of this scene to improve its overall composition.
2	Shift the camera to make the scene more visually appealing.
3	Adjust the framing by shifting to create a stronger composition.
4	Reframe by shifting to enhance the visual composition.
5	Shift the view to improve the image composition.
6	Reframe by shifting to make the scene more pleasing.
7	Refine the image composition by shifting the frame.
8	Shift the scene to enhance the composition.
9	Enhance the scene by shifting the camera framing.
10	Shift the frame to create a stronger composition.
#	Zoom-in Prompt
1	Zoom in the framing of this scene to improve its overall composition.
2	Tighten the framing by zooming in to make the image more appealing.
3	Zoom in to refine the composition and enhance the visual effect.
4	Narrow the framing by zooming in for a more pleasing result.
5	Adjust the composition by zooming in to strengthen the image.
6	Reframe by zooming in to make the scene look more polished.
7	Refine the composition by zooming in to simplify the frame.
8	Enhance the image by narrowing the view with a zoom-in.
9	Strengthen the composition by zooming in.
10	Make the image look cleaner by zooming in on the framing.
#	View Change Prompt
1	Change the viewpoint of this scene to improve its overall composition.
2	Make a view change to find a better shooting point and improve the composition.
3	Perform a view change to choose a more appealing shooting spot.
4	Select a new shooting view to create a better composition.
5	Change the shooting view to achieve a more visually pleasing shooting point.
6	Explore a different shooting spot to improve the overall composition.
7	Change the shooting view to create a more effective composition.
8	Make a view change to enhance the composition of the scene framing.
9	Apply a view change for a more engaging and visually strong composition.
10	Change the viewpoint to make the composition more attractive.

Table A8. **Task prompt** of two auto tasks.

#	Prompt for Static Auto Task (Shift or Zoom-in)
1	Refine the composition through shift or zoom-in adjustments.
2	Improve the composition by combining shift or zoom-in operations.
3	Enhance the image composition with coordinated shift or zoom-in refinement.
4	Adjust the framing through shift or zoom-in to achieve a better composition.
5	Make the composition more appealing using shift or zoom-in adjustments.
6	Refine the scene composition with gentle shift or zoom-in movement.
7	Improve the framing by applying continuous shift or zoom-in optimization.
8	Enhance the overall composition through integrated shift or zoom-in refinement.
9	Adjust the frame smoothly using shift or zoom-in to strengthen composition.
10	Refine the image framing through natural shift or zoom-in enhancement.
#	Prompt for Full Auto Task (All Three Tasks)
1	Capture this scene with better composition.
2	Enhance the composition of this scene.
3	Refine the composition to make this scene more visually pleasing.
4	Adjust the image to improve its composition.
5	Reframe this scene to achieve a more pleasing visual composition.
6	Refine the framing through shift, zoom-in, or view change.
7	Improve the composition through shift, zoom-in, or viewpoint adjustment.
8	Enhance the image composition through shift, zoom-in, and view change.
9	Optimize the composition through shift, zoom-in, or new viewpoint exploration.
10	Refine the scene composition through shift, zoom-in, or view change as needed.

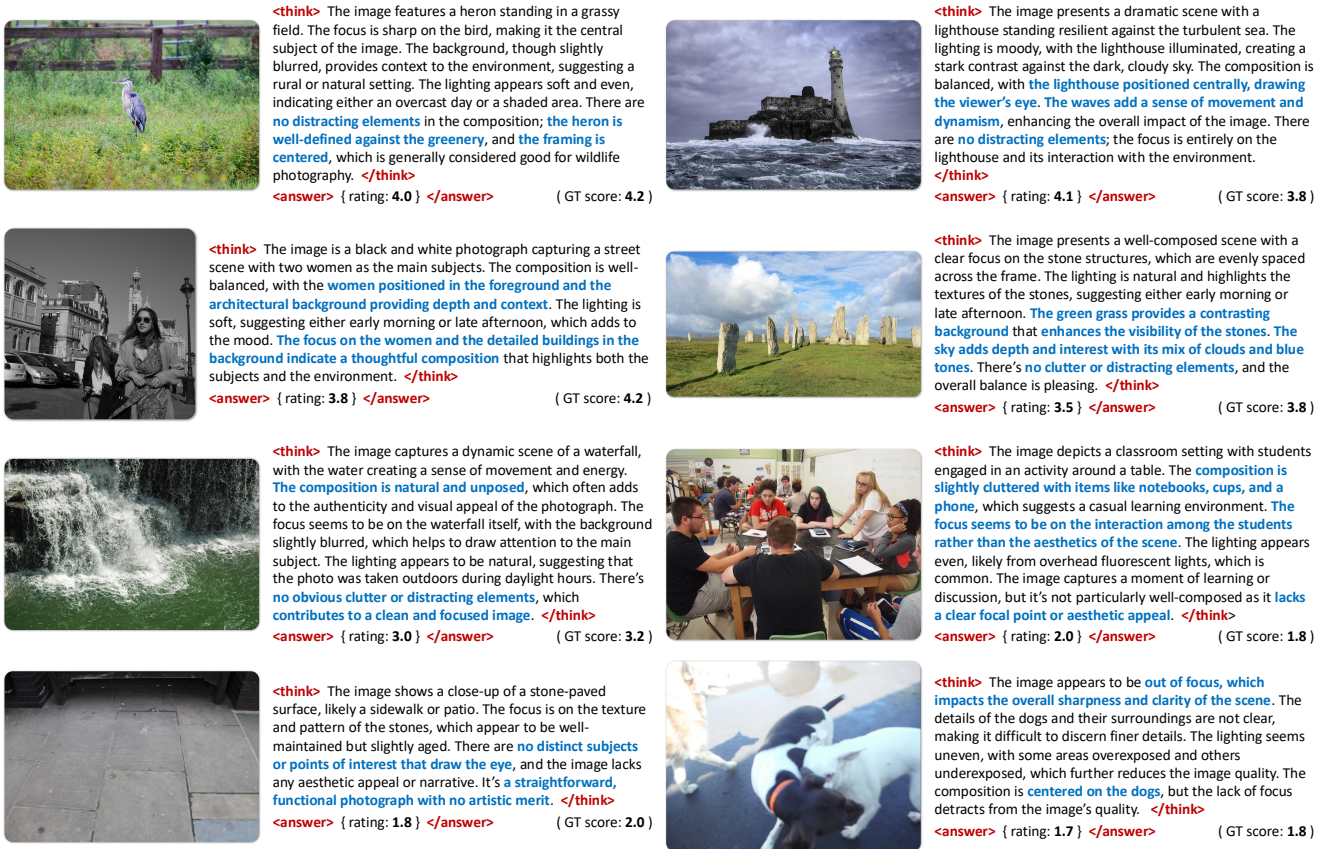


Figure A9. Qualitative results of image composition assessment on CADB dataset.



Figure A10. Qualitative results of image composition classification on CADB dataset.



To enhance the composition, adjust the framing by **cropping out the lower portion of the field, focusing on the upper half where the white structure stands prominently against the tree line**. This change emphasizes the contrast between the bright field and the darker trees, creating a stronger visual hierarchy. Additionally, **include more of the sky above the trees** to introduce a sense of openness and balance the vertical elements. This modification draws attention to the structure as the central subject, while the remaining greenery provides a natural frame, improving overall harmony and aesthetic appeal.



To enhance the composition, crop the image to **focus on the person in the red coat as the central subject**, placing them slightly off-center to adhere to the rule of thirds. **Remove the top portion of the image where the sky dominates**, as it currently draws attention away from the main subject. **Adjust the bottom edge to include more of the snowy path leading into the distance**, creating a stronger leading line that guides the viewer's eye through the scene. This modification will emphasize the subject while maintaining the rural winter setting, improving overall balance and visual flow.



To enhance the composition, adjust the framing by **cropping out the lower portion of the image, focusing on the upper half where the silhouette of the boat and the sunset are more prominent**. This change eliminates unnecessary water surface details, creating a cleaner and more balanced visual weight. By **centering the boat horizontally and aligning it with the reflection of the sun**, the composition becomes more symmetrical and aesthetically pleasing. The reduced water area draws attention to the warm hues of the sunset and the silhouette, emphasizing the serene and tranquil mood of the scene. This modification ensures the viewer's focus remains on the key elements without distraction.



To enhance the composition, **crop the image to focus on the colorful birdhouse and its reflection in the water**, placing them closer to the center. **Remove the distracting empty branches on the left side** to create a cleaner frame. Adjust the background by reducing the visible area of the bridge and distant trees, emphasizing the serene water and the birdhouse as the primary subjects. This shift draws attention to the vibrant colors and reflections, creating a more balanced and aesthetically pleasing image. The tighter framing also eliminates unnecessary elements, making the scene feel more intimate and focused.



To enhance the composition, **crop the image slightly closer to the subjects' faces**, ensuring **they are centered and more prominent**. Remove any unnecessary background elements that distract from the focus on the individuals. **Adjust the framing so that the lights in the background create a soft bokeh effect**, adding depth without overpowering the subjects. This tighter crop will emphasize their expressions and attire, making them the clear focal point while maintaining a balanced and aesthetically pleasing composition.



To enhance the composition, crop the image slightly **closer to the lighthouse and house, focusing on their structures and the rocky foreground**. **Remove the excess empty space above the clouds** to emphasize the dramatic sky and create a stronger visual connection between the elements. This adjustment draws attention to the lighthouse as the central subject, while the house adds depth and context. The tighter framing also balances the weight of the sky and sea, making the overall scene more dynamic and visually engaging.



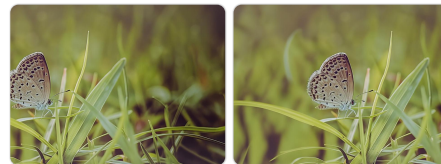
To enhance the composition, crop the image to **focus on the central area where the tram tracks lead into the frame**, creating a strong leading line that draws the viewer's eye deeper into the scene. **Remove the lower portion of the snowy sidewalk**, which appears less dynamic and reduces visual clutter. **Adjust the framing to include more of the illuminated building and its festive decorations**, such as **the "Cafe" sign** and the Christmas tree, to emphasize the warm, inviting atmosphere. This shift balances the composition by highlighting the architectural details and the activity around them, making the scene feel more vibrant and engaging.



To enhance the composition of the original image, adjust the framing by **lowering the horizon line to include more of the foreground**, such as **the road and any visible structures or vegetation**. This change will create a stronger sense of depth and context, grounding the majestic mountains in their environment. Additionally, **ensure that the road leads the viewer's eye towards the mountains**, using the natural lines of the landscape to guide attention. This modification balances the vastness of the sky with the earthy tones of the foreground, making the overall scene more dynamic and engaging.



To enhance the composition, adjust the framing by **centering both dogs more prominently in the scene**. **Ensure they are fully visible**, with their entire bodies including their heads and legs, to emphasize their playful energy. **Remove any distracting elements on the left side**, such as **the partial view of the dog**, to create a cleaner focus on the two main subjects. Additionally, slightly increase the depth of field to **keep the background softly blurred**, ensuring the dogs remain the focal point. This adjustment balances the composition, highlights the dynamic movement of the dogs, and creates a more engaging and aesthetically pleasing image.



To enhance the composition of the original image, reframe it by **centering the butterfly more prominently** within the frame. This adjustment ensures the butterfly is the clear focal point, drawing attention to its intricate wing patterns and delicate details. Additionally, **crop out any unnecessary grass blades on the left side** to eliminate distractions and create a cleaner, more balanced composition. By doing so, the greenery serves as a soft, natural backdrop that complements the butterfly without overpowering it, resulting in a more aesthetically pleasing and focused image.

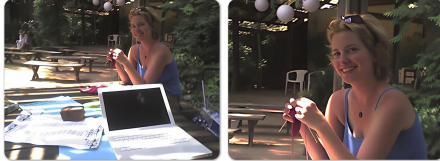
Figure A11. Qualitative results of the shift task. Left / right / bottom: original image / generated example image / text guidance. Actionable textual suggestions are highlighted in **blue**, while hallucinated descriptions are marked in **red**.



To enhance the composition, **zoom in on the skier**, centering them more prominently. **Remove the top portion with the banner and people in the background** to minimize distractions. **Position the skier slightly off-center** to follow the rule of thirds. This tighter framing focuses attention on the skier's action and the snowy slope, creating a dynamic and engaging image.



To enhance the composition, zoom in on the surfer and wave, **removing the top-left tree branches and bottom-right observer**. This tight framing focuses attention on the surfer's action and the wave's texture. **Centering the surfer within the wave** adheres to the rule of thirds, creating balance. Eliminating distractions sharpens the viewer's focus, resulting in a more dynamic and engaging image.



To improve the composition, **zoom in closer to the woman, focusing on her face and upper body** while **excluding the laptop and knitting materials**. This shift centers attention on her expression and interaction with the knitting needles, creating a more intimate and engaging scene. By removing the laptop, you eliminate distractions and emphasize the personal, creative activity. Adjust the framing to **place her slightly off-center**, following the rule of thirds, which adds visual interest and balance. This tighter crop also simplifies the background, drawing the viewer's eye to her actions and emotions.



To enhance the composition of the image, zoom in and reframe to **focus on the bus as the central subject, eliminating unnecessary foreground elements like the person walking and the parked cars on the left**. This adjustment reduces distractions and directs attention to the bus's vibrant colors and movement. **Shift the framing slightly to place the bus near the center** but not perfectly aligned, adhering to the rule of thirds for a more dynamic feel. By removing the lower portion of the image, including the road markings and some vehicles, the composition becomes cleaner and more focused, emphasizing the bus's journey against the backdrop of the large building and flags. This reframing creates a stronger visual narrative and balance.



To enhance the composition, zoom in on the subject by **cropping out the lower portion** of the image, focusing on the man from just above his waist up. This adjustment **eliminates unnecessary ground space**, creating a more dynamic and centered composition. By keeping the subject slightly off-center (to the left), you adhere to the **rule of thirds**, adding visual interest. The tight framing also removes extraneous details in the background, such as the lower part of the building and the floor, which helps draw attention to the subject's gesture and expression. The vibrant purple lighting behind him remains prominent but is now framed more intentionally, enhancing the overall aesthetic balance and focus.



To enhance the composition, **zoom in on the tennis player**, centering him more prominently. This **eliminates excess background and focuses attention on his action**. Adjust the angle slightly to align him with the court lines, **adhering to the rule of thirds**. This tighter framing creates a dynamic, engaging image that emphasizes the athlete's movement and the vibrant court colors.



To improve the composition of the image, zoom in and crop the scene to **focus more tightly on the cat and the ornate green bench**. **Eliminate the excess yellow wall space above and below the bench**, which currently dominates the frame. By reframing, the cat becomes the central subject, drawing immediate attention due to its position and the contrast against the vibrant green background. This tighter crop also accentuates the intricate details of the bench's design, creating a more dynamic and engaging visual. Additionally, removing the empty space around the edges reduces distractions and emphasizes the harmony between the cat and its environment, adhering to the rule of thirds for a balanced and aesthetically pleasing result.



To enhance the composition of the image, zoom in and reframe to **focus on the narrow gauge train** as the central subject, **eliminating unnecessary foreground elements like the road and the distant house**. This adjustment brings attention to the train's details and its interaction with the landscape, such as the vibrant orange flowers and rolling hills. By placing the train slightly off-center, following the rule of thirds, the composition becomes more dynamic and balanced. The tighter frame also reduces distractions from peripheral elements, emphasizing the harmony between the train and the scenic backdrop, creating a more cohesive and aesthetically pleasing image.



To enhance the composition, zoom in slightly to **center the train more prominently**. This adjustment **reduces empty space on the sides, focusing attention on the train and its lights**. Position the train off-center following the rule of thirds for a balanced look. Eliminate distractions like distant poles and shrubs to emphasize the train and the mountainous backdrop, creating a clearer, more engaging image.



To enhance the composition, zoom in and reframe to **focus on the lighthouse and its immediate surroundings**. **Crop out the top portion of the sky**, which is mostly empty and does not add significant visual interest. **Shift the framing slightly to bring the lighthouse and rocky outcrop closer to the center**, emphasizing their prominence. Reduce the visible expanse of the ocean on the left side to create a more balanced and intimate view. This reframing brings attention to the interplay between the natural landscape and the man-made structure, enhancing the aesthetic by creating a more dynamic and focused scene.

Figure A12. Qualitative results of the zoom-in task. Left / right / bottom: original image / generated example image / text guidance. Actionable textual suggestions are highlighted in blue, while hallucinated descriptions are marked in red.



To enhance the composition, adjust the viewpoint to **emphasize the expansive coastal landscape while reducing the prominence of the dry, thorny foreground vegetation**. **Shift focus slightly upward to reveal more of the serene ocean horizon and the distant coastline**, creating a sense of depth and vastness. **Crop out some of the distracting, dense vegetation in the lower portion**, allowing the natural beauty of the flowers and the open sky to take center stage. This change balances the elements, highlighting the contrast between the rugged terrain and the tranquil sea, resulting in a more harmonious and aesthetically pleasing image.



To enhance the composition, adjust the viewpoint to **focus more on the vibrant yellow trees**, ensuring they occupy a larger portion of the frame. **Crop out the lower foreground rocks and dirt path** to minimize distractions and emphasize the verticality of the trees. **Shift the horizon line slightly higher to include more of the clear blue sky**, which will balance the warm tones of the foliage and add depth. This change will highlight the contrast between the golden trees and the cool sky, creating a more striking and aesthetically pleasing image.



To enhance the composition, **zoom in on the sunflower and crop out the background elements such as the road, trees, and distant structures**. This focuses attention on the flower and its immediate surroundings, creating a more intimate and striking visual. By **removing distractions like the empty parking lot and utility poles**, the viewer's gaze is drawn directly to the vibrant yellow sunflower, which becomes the central point of interest. Additionally, this closer framing emphasizes the contrast between the delicate flower and the rough texture of the surrounding rocks, enhancing the overall aesthetic appeal and emotional impact.



To enhance the composition, adjust the viewpoint to **place the dome of the building more centrally and prominently** within the frame, ensuring it becomes the focal point. **Remove the foreground trees that partially obscure the structure**, allowing the full grandeur of the architecture to be appreciated. **Shift the angle slightly upward to emphasize the height and symmetry of the building against the clear blue sky**. This change will create a more balanced and aesthetically pleasing image, highlighting the historical and architectural significance of the scene. Additionally, consider reducing distractions such as the yellow structure on the right side to maintain visual harmony.



To enhance the composition, reposition the viewpoint to **include more of the expansive ocean and distant horizon**, creating a sense of scale and depth. **Shift focus slightly upward to minimize the dominance of the sandy foreground and emphasize the rugged textures of the cliffs**. Incorporate more sky to **balance the heavy rock formations and add visual interest with clouds or clear blue tones**. This adjustment will draw the viewer's eye across the scene, from the detailed cliff textures to the vastness of the sea and horizon, resulting in a more dynamic and aesthetically pleasing image.



To enhance the composition, adjust the viewpoint by **focusing on a single prominent leaf in the foreground**, creating a clear point of interest. **Shift the perspective slightly upward to minimize the dominance of the asphalt surface** and emphasize the scattered leaves and their natural patterns. **Include more of the background trees and foliage** to add depth and context, **ensuring the sun's glow is visible through the leaves** to introduce warmth and contrast. This change directs the viewer's gaze naturally along the path, enhancing the sense of movement and seasonal ambiance.



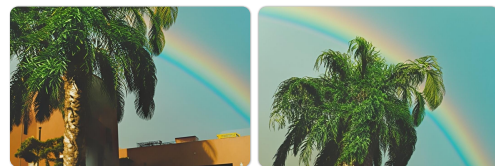
To enhance the composition, adjust the viewpoint to **focus on the dramatic interplay between the clouds and the landscape**. Place the horizon line higher in the frame to **emphasize the expansive sky and cloud cover**, creating a sense of grandeur and scale. **Remove the signpost from the right side to eliminate distractions** and allow the natural elements to take center stage. **Shift slightly to include more of the tree line at the edge of the cloud layer**, which will add depth and visual interest. This change directs attention to the contrast between the earthy tones below and the ethereal quality of the clouds above, resulting in a more balanced and aesthetically pleasing image.



To enhance the composition, adjust the viewpoint to **focus more on the expansive green field and the distant cityscape**, creating a stronger sense of depth and scale. Place the horizon line **higher** in the frame to **emphasize the vastness of the open space and the urban skyline beyond**. **Remove distractions such as the fence and people near the bottom**, allowing the viewer's eye to flow freely across the field. Incorporate more of the palm trees and surrounding buildings to add texture and context without overcrowding the scene. This shift will balance the natural and urban elements, highlighting the contrast between the serene park and the bustling city.



To enhance the composition, reposition the viewpoint to **focus solely on the urban skyline at night**, eliminating the foreground vegetation and pole. This simplifies the scene, allowing the illuminated buildings to become the central visual element. By **removing distractions such as the natural elements and the pole**, the viewer's attention is drawn directly to the architectural details and the interplay of light and shadow. The resulting image will emphasize symmetry, depth, and the vibrant cityscape, creating a more cohesive and aesthetically pleasing photograph.



To enhance the composition, **focus solely on the palm tree and rainbow**, removing the building from the frame. **Place the tree centrally** to emphasize its natural elegance and **allow the rainbow to arc gracefully across the clear sky**, creating a strong visual contrast. **Eliminate distractions such as the balcony and roofline**, which currently compete for attention. This adjustment will highlight the serene beauty of the scene, balancing the vibrant colors of the rainbow against the lush green foliage and the soft blue sky. The result will be a more harmonious and aesthetically pleasing image that draws the viewer's eye naturally to the interplay between nature's elements.

Figure A13. Qualitative results of the view-change task. Left / right / bottom: original image / generated example image / text guidance. Actionable textual suggestions are highlighted in blue, while hallucinated descriptions are marked in red.