Venus: Benchmarking and Empowering Multimodal Large Language Models for Aesthetic Guidance and Cropping

Tianxiang Du, Hulingxiao He, Yuxin Peng *
Wangxuan Institute of Computer Technology, Peking University

arXiv preprint arXiv:2602.23980, 2026.

Motivation & Background

- The ubiquity of smartphone photography has made taking photos effortless, yet a significant gap remains between ordinary users and professional photographers.
- Professional photographers can identify aesthetic issues (e.g., imbalanced composition, improper lighting) and provide actionable shooting suggestions.
- The authors define this capability as **Aesthetic Guidance** (AG).

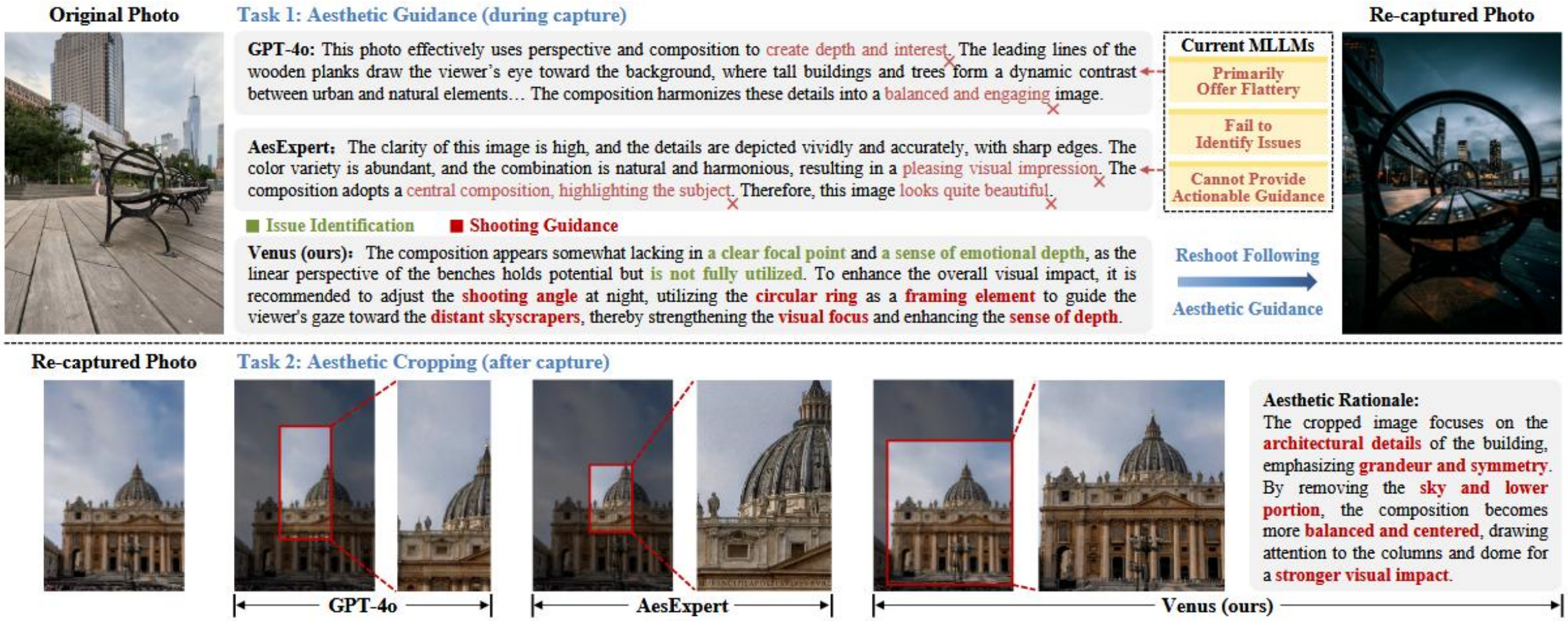
Original Photo

Reshoot Following

Aesthetic Guidance

Re-captured Photo

Motivation & Background



- Existing multimodal large language models (MLLMs) primarily offer **overly positive** feedback, failing to **identify issues** or provide **actionable guidance**.

Related Work

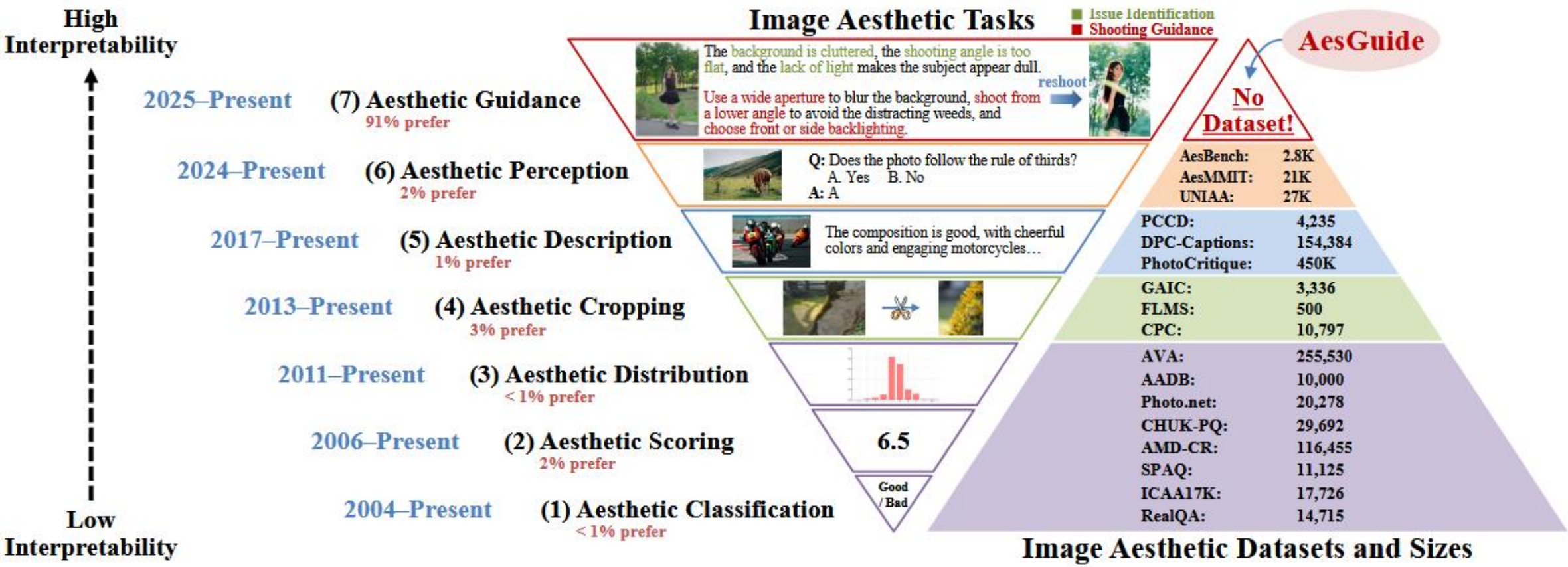


Figure 2. Overview of image aesthetic tasks and datasets. We follow and refine the comprehensive aesthetic task taxonomy proposed by Jin *et al*. [20]. A user survey of 1,069 participants shows that 91% prefer AG, a largely underexplored task with no dedicated dataset.

Contribution

- We formally define **aesthetic guidance (AG)**, an underexplored domain bridging subjective aesthetic understanding with objective, actionable shooting adjustments.
- **AesGuide** is presented as the first AG benchmark, annotated with aesthetic scores, analyses, and guidance, along with three evaluation metrics and expert assessments.
- We propose **Venus**, a two-stage framework that first empowers MLLMs with AG capability and then unlocks their aesthetic cropping power via CoT-based rationales.
- Extensive experiments show that Venus achieves SOTA performance in both AG and aesthetic cropping, enabling **interpretable** and **interactive** aesthetic refinement.

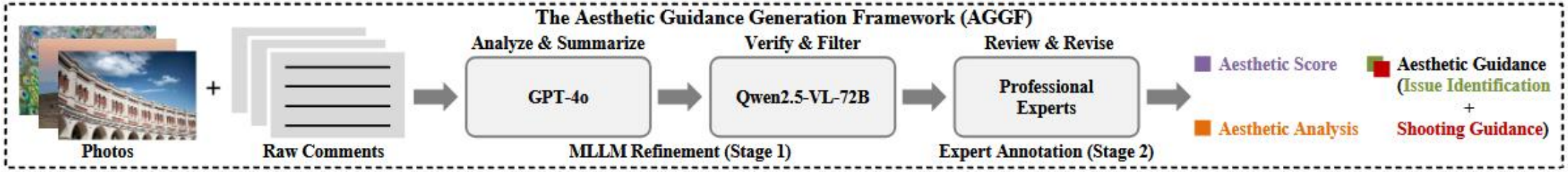
AesGuide Dataset

- It contains 10,748 real-world photos, each annotated with an aesthetic score, analysis, and guidance.
- The dataset is collected from two main sources: online platforms and professional photographers.
- However, these raw comments are often noisy and inconsistent, making them unsuitable for direct supervised fine-tuning.

- Benchmark: 1,000 images + GPT-4o automatic evaluation

Prompt: *Evaluate whether [RESPONSE] contains the image's shortcomings or suggestions for improving the image mentioned in [ANSWER].*

1. *If [RESPONSE] does not contain the image's shortcomings or suggestions for improving the image mentioned in [ANSWER], rate a score of 0.*
2. *If [RESPONSE] contains some of the image's shortcomings or suggestions mentioned in [ANSWER], rate a score of 1.*
3. *If [RESPONSE] contains most of the image's shortcomings or suggestions mentioned in [ANSWER], rate a score of 2.*



AesGuide Dataset

Aesthetic Score: 1.4

This photo exhibits high visual clarity, with a bright blue sky and clouds that appear vivid and dynamic.

The composition lacks refinement. The utility pole and diagonal wires introduce visual clutter. It is recommended to change the shooting angle to avoid these distracting elements and better highlight the clouds and sky. Alternatively, one could wait for an interesting interaction between the pole and the clouds, using the pole as a compositional element to add narrative depth to the image.

Aesthetic Score: 5.2

The atmosphere in this photo is captivating, with the water reflections and the ripples left by the swans adding both depth and movement to the scene. The graceful posture of the two black swans infuses the composition with vitality and energy.

The two swans are positioned at opposite edges of the frame, leaving the center empty. Capturing them when they are closer together, adding additional elements to fill the space, or waiting for an interesting interaction between them would create a more cohesive and visually engaging composition.

Aesthetic Score: 8.9

The idea and composition are excellent. The contrast between the foreground traditional architecture and the background modern skyscrapers emphasizes the clash of time and culture, creating a powerful visual impact. The use of light enriches the narrative.

The roof of the traditional building is slightly overexposed, causing detail loss, and the cluttered branches in the foreground disrupt the atmosphere. Adjusting exposure or using a gradient filter can help preserve detail, while avoiding the branches would improve the cleanliness of the composition.

Aesthetic Score: 3.2

The idea of this photo is excellent, with the children using a spotlight to illuminate a sand pit on the beach, creating a strong sense of narrative.

The main subject is overly concentrated in the bottom left corner, and the overall image appears quite dark. It is recommended to increase the negative space on the left side to shift the visual focus (the children) towards the center of the frame. Additionally, slightly increasing the exposure can help brighten the image and enhance its vibrancy.

Aesthetic Score: 7.7

This image excels in emotional expression, with light and shadow enhancing the three-dimensionality of the scene. The desert symbolizes solitude and vastness, while the camel's movement and the person's persistence convey resilience and exploration.

The background feels somewhat monotonous, lacking contrast or a connection with the main subject. Employing a wide aperture to blur the background or incorporating natural elements, such as the curves of sand dunes or distant landscapes, could highlight the desert's vastness and enhance the overall depth.

Aesthetic Score: 9.4

It utilizes symmetry and horizontal composition to create a balanced effect. The reflections of the figures in the water add visual interest, while the calm background contrasts with the dynamic movement of the people, enhancing the artistic appeal.

The image is slightly tilted, and the central figure appears a bit crowded. It is recommended to adjust the camera angle to ensure the lines are parallel, enhancing the symmetry and balance. Additionally, reducing the number of figures could simplify the scene and enhance the artistic effect.

Overview of the Venus Framework

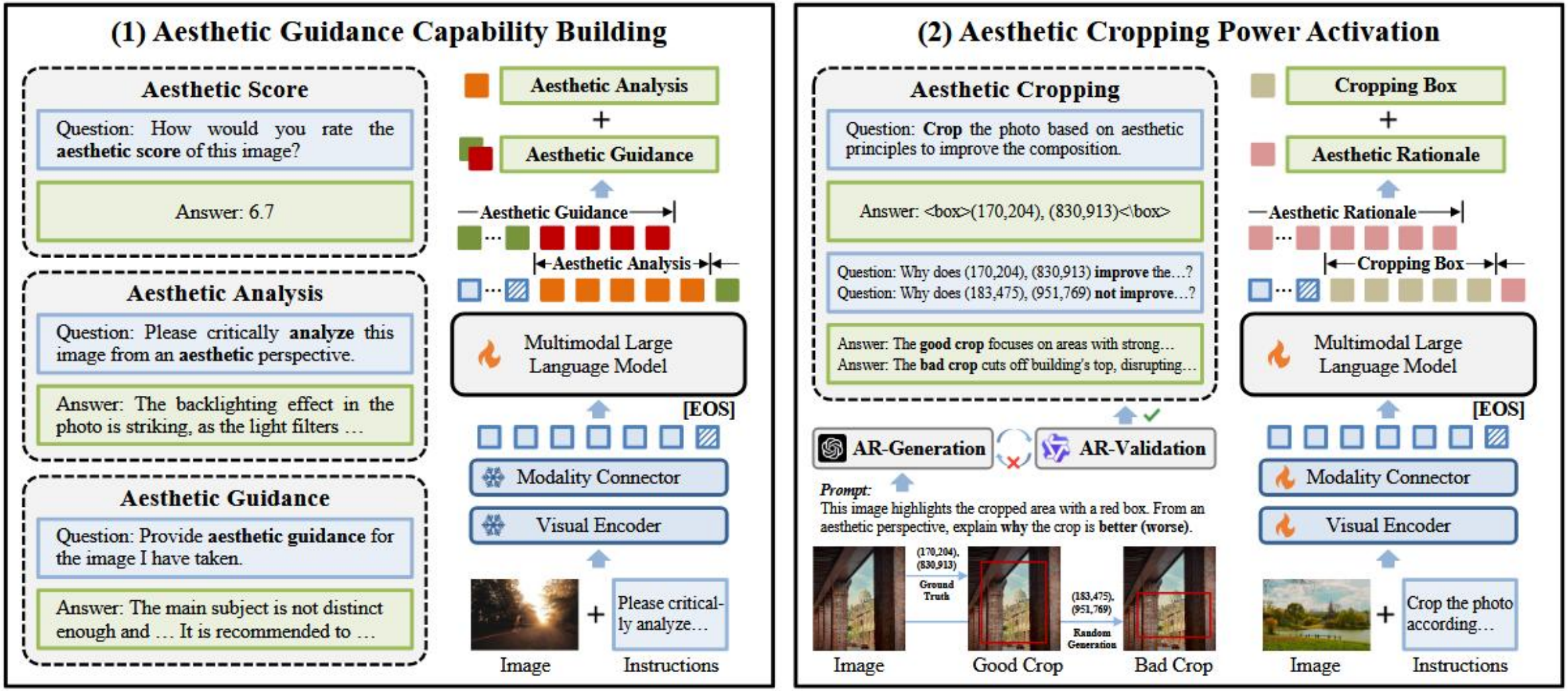
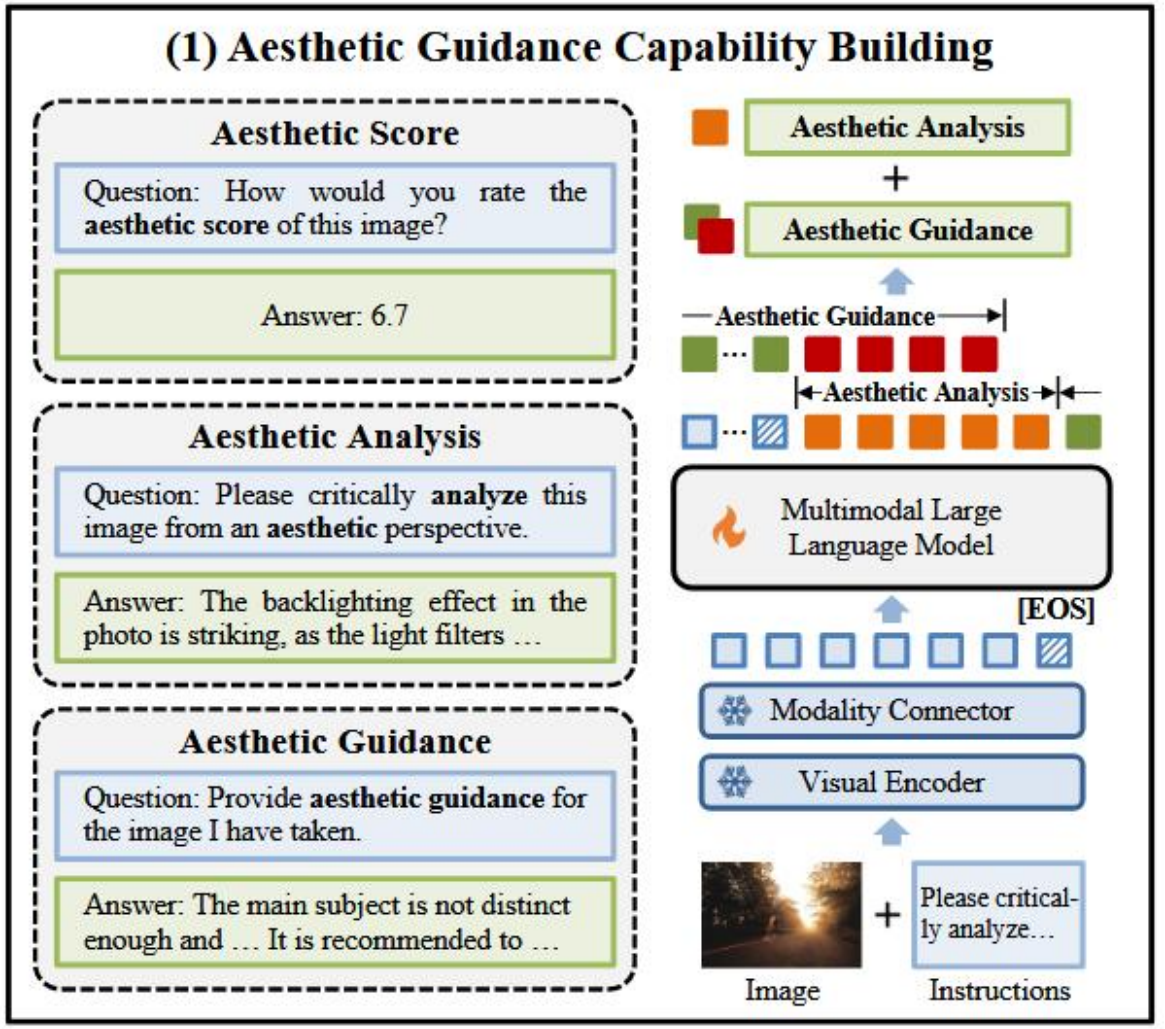


Figure 4. Overview of the **Venus** framework: (1) Aesthetic guidance capability building, where AesGuide is leveraged to empower MLLMs with AG capability. (2) Aesthetic cropping power activation, which unlocks the cropping ability using CoT-based rationales.

Aesthetic Guidance Capability Building

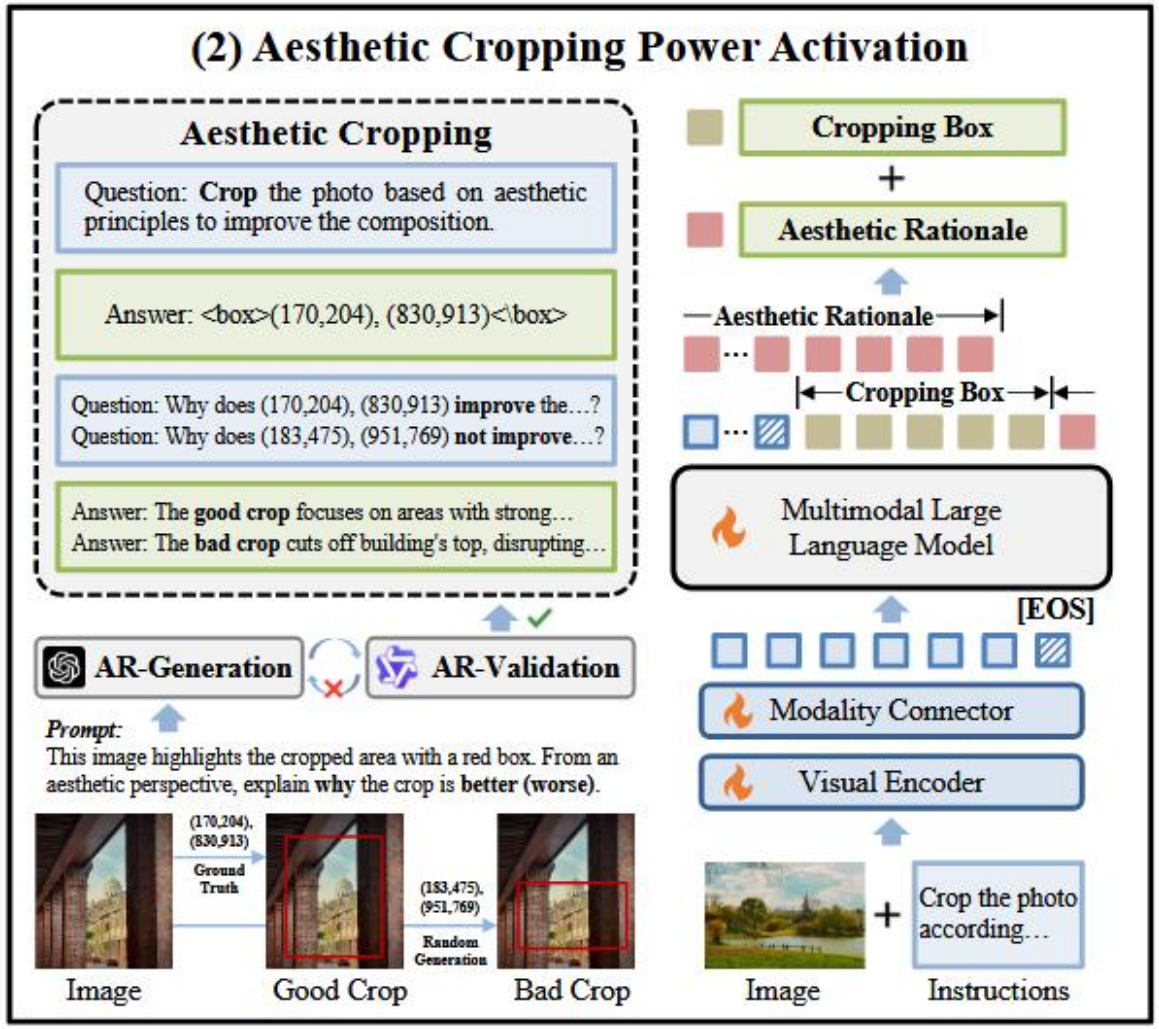
- We apply **supervised instruction fine-tuning** on the AesGuide dataset to equip the MLLM with AG capabilities.
- During this process, the visual encoder and modality connector are frozen, while only the LLM is updated.

$$\mathcal{L}_{AG} = -\mathbb{E}_{(x,q,a,g)\sim\mathcal{AG}} \sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, q, y_{<t}),$$



Aesthetic Cropping Power Activation

- We activate the aesthetic cropping power of the aesthetic guidance MLLM using highquality CoT-based rationales.
- **AR-Generation.** GPT-4o is prompted with the original image, where the cropping region is outlined in red, and instructed to explain why the area inside the box exhibits better or worse composition.
- **AR-Validation.** The generated rationales are then reviewed by Qwen2.5-VL-72B, which verifies whether each explanation aligns with the visual content and correctly evaluates the cropping quality.



Experiment

Table 1. Evaluation results on the AesGuide benchmark. **Mean** denotes the average across three dimensions. **Expert** represents expert assessments, and **Rank** shows the order by Mean / Expert scores. Models are grouped by category: proprietary, aesthetic-oriented, and open-source general MLLMs. Best and second-best results are in **bold** and underlined, respectively.

Model	Params	Completeness	Preciseness	Relevance	Mean	Expert	Rank
Proprietary MLLMs							
GPT-4o [16]	NA	0.84	1.09	1.01	0.98	1.15	8 / 7
Gemini-2.0-Pro [38]	NA	1.09	1.12	1.36	1.19	1.16	6 / 6
Qwen-VL-Max [2]	NA	0.90	1.05	0.56	0.84	0.89	11 / 11
Aesthetic MLLMs							
AesExpert [14]	7B	0.33	0.56	0.51	0.47	0.56	15 / 14
UNIAA [57]	7B	1.03	1.02	1.23	1.09	1.01	7 / 8
Open-source General MLLMs							
Qwen-VL-Chat [3]	7B	0.73	0.91	0.59	0.74	0.70	12 / 12
Venus-Q (ours)	7B	1.12 +0.39	1.23 +0.32	1.57 +0.98	1.31 +0.57	1.36 +0.66	5 / 4
InternVL 2.5 [8]	7B	0.83	1.01	1.02	0.95	0.99	10 / 9
Venus-I (ours)	7B	1.27 +0.44	1.33 +0.32	1.81 +0.79	1.47 +0.52	1.50 +0.51	2 / 2
MiniCPM-V 2.6 [48]	7B	0.83	1.04	1.04	0.97	0.92	9 / 10
Venus-M (ours)	7B	1.19 +0.36	1.24 +0.20	1.72 +0.68	1.38 +0.41	1.30 +0.38	4 / 5
LLaVA-1.5-7B [26]	7B	0.64	0.79	0.35	0.59	0.52	14 / 15
Venus-L-7B (ours)	7B	1.26 +0.62	1.32 +0.53	1.80 +1.45	1.46 +0.87	1.40 +0.88	3 / 3
LLaVA-1.5-13B [26]	13B	0.67	0.86	0.41	0.65	0.61	13 / 13
Venus-L-13B (ours)	13B	**1.28** +0.61	**1.35** +0.49	**1.83** +1.42	**1.49** +0.84	**1.53** +0.92	**1 / 1**

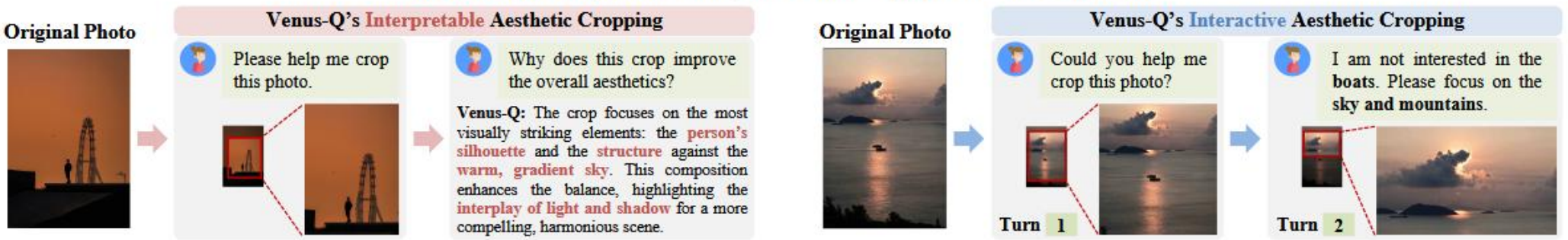
Experiment

Model	IoU%(↑)	Disp(↓)	R%(↑)	Inp.	Ina.
Specialized Aesthetic Cropping Models					
ASM-Net [39]	84.86	0.0390	-	✘	✘
CACNet [11]	85.40	0.0330	-	✘	✘
HCIC [52]	85.00	0.0340	-	✘	✘
SAC-Net [45]	85.51	0.0333	-	✘	✘
UNIC [27]	84.00	0.0370	-	✘	✘
ProCrop [53]	84.30	0.0360	-	✘	✘
Proprietary MLLMs					
GPT-4o [16]	71.61	0.0615	43.2	✓	✓
Gemini-2.0-Pro [38]	68.15	0.0647	38.4	✓	✓
Qwen-VL-Max [2]	58.09	0.0814	20.4	✓	✓

Model	IoU%(↑)	Disp(↓)	R%(↑)	Inp.	Ina.
Aesthetic MLLMs					
AesExpert [14]	37.86	0.1206	9.6	✓	✓
UNIAA [57]	21.03	0.1602	0.2	✓	✓
Open-source General MLLMs					
Qwen-VL-Chat [3]	73.84	0.0667	67.2	✓	✓
Qwen2.5-VL-7B [4]	50.32	0.1056	23.6	✓	✓
Qwen2.5-VL-32B [4]	54.35	0.0957	19.0	✓	✓
InternVL 2.5 [8]	71.53	0.0705	51.2	✓	✓
MiniCPM-V 2.5 [48]	49.29	0.1158	12.5	✓	✓
MiniCPM-V 2.6 [48]	45.95	0.1111	5.2	✓	✓
LLaVA-1.5-7B [26]	54.30	0.0793	21.6	✓	✓
LLaVA-1.5-13B [26]	59.04	0.0876	32.0	✓	✓
Venus-Q (ours)	**87.01** **+1.50**	**0.0292** **-0.0038**	**92.0** **+24.8**	✓	✓

Experiment

(a) Qualitative comparison among GPT-4o, AesExpert, and Venus-Q (ours)



(b) Demonstration of Venus-Q's interpretable and interactive aesthetic cropping

Ablation Study

Table 3. Ablation study on the AesGuide benchmark.

Settings	Com.	Pre.	Rel.	Mean
w/o Aesthetic Analysis in Stage 1	1.09	1.17	1.41	1.22
Venus-Q (ours)	**1.12**	**1.23**	**1.57**	**1.31**

Table 4. Ablation study on the FLMS benchmark.

Settings	IoU%(↑)	Disp(↓)	R%(↑)
w/o AesGuide in Stage 1	84.90	0.0327	86.2
w/o AR-Generation in Stage 2	81.10	0.0402	75.2
w/o AR-Validation in Stage 2	86.47	0.0308	90.4
Venus-Q (ours)	**87.01**	**0.0292**	**92.0**

Thanks