LEARNING AESTHETIC PREFERENCES IN CONTEXTUALLY RELATED IMAGE COLLECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid advancement of smart devices and social media has significantly increased the user demand for aesthetic applications such as album management, photo sharing, and recommendation. In these real-world scenarios, users generally select preferred photographs from a collection of images depicting the same scene, event, or subject, making aesthetic preference inherently context-dependent. However, existing aesthetic evaluation methods typically focus on either context-independently or highly similar image series, limiting their ability to capture preferences within contextually related image collections. To address this gap, we introduce **ContextAes-40K**, a novel dataset comprising 39,033 images organized into 8,069 contextually related collections, with pairwise aesthetic preference annotations within each collection. To further evaluate performance in real-world user-curated collections, we collect **ContextAes-Album**, an independent real-user album dataset with user-provided aesthetic rankings. Building upon ContextAes-40K, we propose **ContextAes**, an MLLM-based aesthetic model trained through a three-stage training strategy. Extensive experimental results show that ContextAes achieves state-of-the-art performance on ContextAes-40K, exhibits strong generalization ability on other aesthetic benchmarks, and also demonstrates superior zero-shot performance on ContextAes-Album, highlighting its ability to identify aesthetic preferences within contextually related image collections.

1 INTRODUCTION

Modern digital photography and social media have transformed how people create, organize, and share visual content, making the selection of aesthetically appealing images an increasingly common task (Loui & Savakis, 2003; Zhao et al., 2025; Sun et al., 2017). Unlike independent image assessment, many such decisions are made within collections of images that share a common context, such as naturally co-occurring images of the same scene or multiple edited variants derived from the same source image. In such settings, aesthetic preference is inherently relative: users often need to determine which images are preferred among a set of related alternatives rather than judge each image in isolation. However, existing aesthetic assessment paradigms typically focus on independent images or highly similar image series, leaving aesthetic preference assessment within contextually related image collections underexplored.

Research on image aesthetic assessment (IAA) has evolved from handcrafted feature-based methods to deep learning and, more recently, multimodal large language model (MLLM)-based approaches. Large-scale datasets such as AVA (Murray et al., 2012) have established a widely adopted paradigm in which each image is assigned an aesthetic quality score, enabling models to learn general aesthetic criteria from diverse visual content. However, such methods typically evaluate images independently, without explicitly considering their relationships within a collection. Subsequent studies have enriched aesthetic modeling by introducing additional forms of structure. For instance, TAD66K (He et al., 2022) organizes images according to semantic themes, but aesthetic assessment is still performed at the individual-image level rather than through within-collection preference comparison. More recently, FGAesthetics (Yang et al., 2026b) has extended its scope to image series, focusing on the aesthetic assessment of highly similar images. While this setting considers relationships among images, it remains limited in scope, with aesthetic assessment in image collections characterized by broader inter-image relationships still underexplored.

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

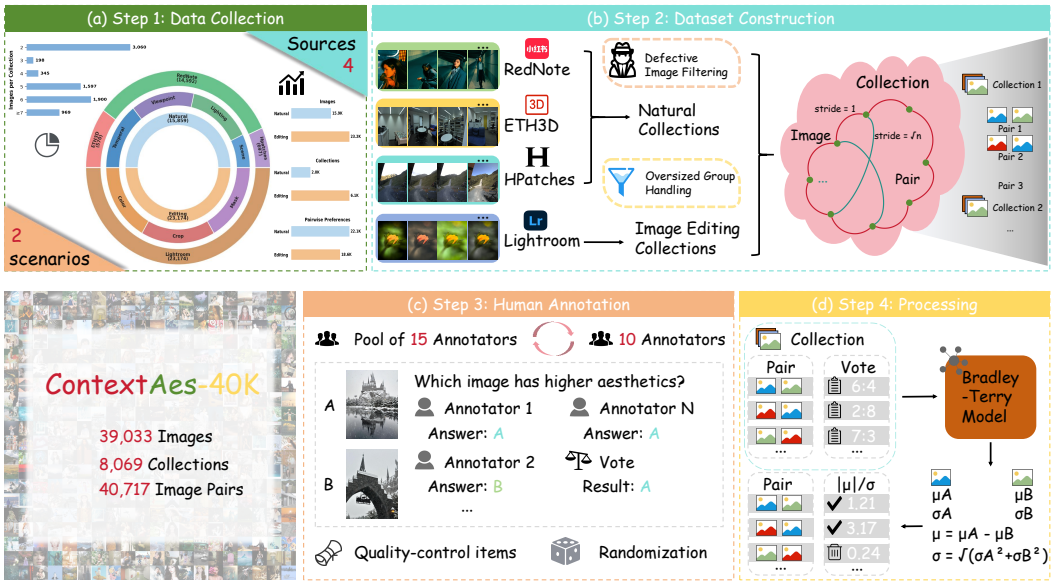


Figure 1: ContextAes-40K construction pipeline. (a) Data Collection. Select images from 4 source datasets covering 2 scenarios. (b) Dataset Construction. Generate pairs within each collection using a regular graph. (c) Human Annotation. Annotators compare each pair to obtain aesthetic preference votes. (d) Processing. Apply the Bradley–Terry model to each collection and remove pairs with high uncertainty or indistinguishable preferences.

To address these challenges, we introduce **ContextAes-40K**, a novel dataset covering contextually related image collections from both natural photography and image editing scenarios. It comprises 39,033 images organized into 8,069 collections, with 40,717 pairwise preference annotations within each collection. We further collect **ContextAes-Album**, an independent real-world dataset of user-curated photo collections with user-provided aesthetic rankings, providing an external evaluation setting for real-world aesthetic preferences.

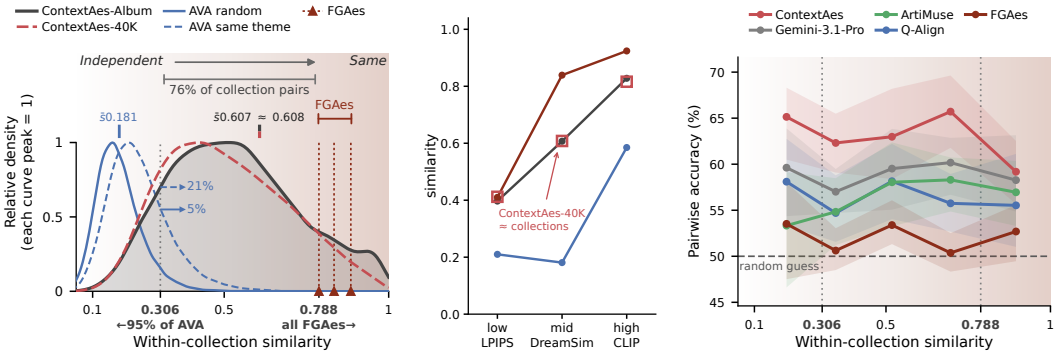
Building upon ContextAes-40K, we further propose **ContextAes**, an MLLM-based aesthetic model for learning aesthetic preferences within contextually related image collections. Rather than directly treating aesthetic preference learning as a standard fine-tuning problem, ContextAes adopts a three-stage training strategy that progressively equips the model with task-relevant aesthetic knowledge. Specifically, SkillOpt provides optimized aesthetic prior knowledge, supervised fine-tuning adapts the model to comparative aesthetic judgment, and GRPO further aligns its predictions with human preference annotations. Extensive experiments demonstrate that ContextAes achieves state-of-the-art performance on ContextAes-40K and exhibits strong generalization across established aesthetic benchmarks. Furthermore, ContextAes outperforms existing methods on the independent real-world ContextAes-Album dataset, providing evidence of its ability to identify aesthetic preferences within contextually related image collections.

The main contributions are summarized as follows:

• We introduce **ContextAes-40K**, a novel dataset for learning aesthetic preferences within contextually related image collections, with 39,033 images organized into 8,069 collections and pairwise aesthetic preference annotations. **ContextAes-Album** further provides an independent real-world evaluation of user aesthetic preferences.

• We propose **ContextAes**, an MLLM-based aesthetic model with a three-stage training strategy that combines aesthetic skill optimization, supervised comparative learning, and GRPO-based preference alignment.

• We conduct extensive experiments demonstrating that ContextAes achieves state-of-the-art performance on ContextAes-40K, exhibits strong generalization across established aesthetic benchmarks, and outperforms existing methods on the real-world ContextAes-Album dataset.



(a) (b) (c)

Figure 2: Analysis of Contextually Related Image Collections. (a) Distributions of ContextAes-40K, ContextAes-Album, AVA, and FGAesthetics based on $1 - \text{DreamSim}$. (b) Distributions measured by LPIPS (low-level), DreamSim (mid-level), and CLIP (high-level) across the four datasets. Note that LPIPS and DreamSim are subtracted from 1 so that higher values indicate greater similarity. (c) Model performance on ContextAes-Album.

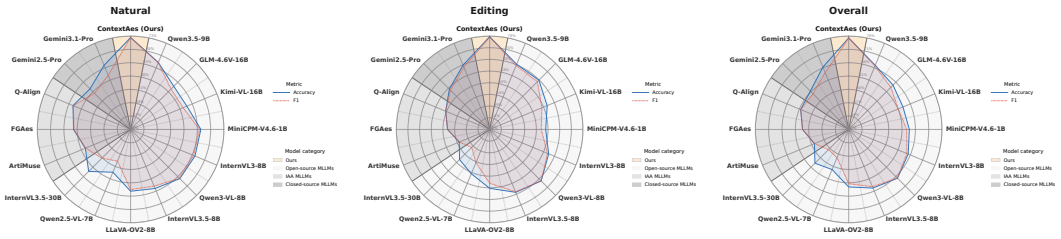
2 Related Work

2.1 Image Aesthetic Assessment Datasets

Existing datasets for image aesthetic assessment support a variety of aesthetic evaluation settings, ranging from holistic judgments of individual images to more specialized forms of aesthetic annotation. AVA (Murray et al., 2012) established a large-scale setting for holistic aesthetic evaluation, while subsequent datasets enriched image-level supervision in different ways. For example, TAD66K (He et al., 2022) organizes images according to semantic themes, while PARA (Yang et al., 2022) introduces rich annotations for personalized image aesthetics. Resources such as AesGuide (Du et al., 2026) and ArtiMuse (Cao et al., 2026) further incorporate richer aesthetic analysis, guidance, and fine-grained annotations. Despite these advances, individual images remain the primary unit of assessment in most existing datasets. More recently, FGAesthetics (Yang et al., 2026b) has extended aesthetic assessment to image series by introducing pairwise aesthetic ranking among highly similar images, marking a shift toward comparative assessment across related images. However, such comparative assessment remains focused on highly similar image series, while aesthetic assessment in collections characterized by broader inter-image relationships remains relatively underexplored. To this end, we introduce ContextAes-40K, a novel dataset covering contextually related image collections from both natural photography and image editing scenarios, with pairwise aesthetic preference annotations within each collection.

2.2 Image Aesthetic Assessment Models

Image aesthetic assessment models have evolved from handcrafted features and conventional machine learning to deep neural networks and multimodal models. Early deep learning approaches such as NIMA (Talebi & Milanfar, 2018) learn aesthetic score distributions directly from images, while subsequent methods have developed stronger visual representations for aesthetic modeling. Vision-language approaches further incorporate textual aesthetic information; for example, VILA (Lin et al., 2024) learns multimodal aesthetic representations from image-comment pairs through vision-language pretraining. More recently, multimodal large language models have been explored for aesthetic perception, description, and scoring. Methods such as Q-Align (Wu et al., 2023) and UNIAA (Zhou et al., 2024) adapt MLLMs to aesthetic assessment through language-guided supervision, while AesExpert (Huang et al., 2024a) and ArtiMuse (Cao et al., 2026) further enhance aesthetic understanding with richer textual analysis and scoring capabilities. Recent studies have also begun to incorporate relationships among images into aesthetic modeling. FGAesQ (Yang et al., 2026b) learns discriminative aesthetic scores from relative ranks within highly similar image series, while RED-Aes (Jia et al., 2026) models aesthetic differences between source images and their edited variants. While existing studies have addressed independent image assessment and narrowly defined forms of image comparison, aesthetic preference assessment within contextually related



(a) (b) (c)

Figure 3: Comparison of Accuracy and F1 Scores of Different Methods on ContextAes-40K across different scenarios: (a) Natural, (b) Editing, and (c) Overall.

image collections has received limited attention. To address this gap, we develop ContextAes, an MLLM-based aesthetic model that demonstrates strong performance in this regard.

3 ContextAes-40K

3.1 Data Construction

To construct ContextAes-40K, we consider two complementary types of contextual relationships: relationships arising from natural photography and those induced by image editing. An overview of the dataset construction pipeline is shown in Figure 1. The former arises naturally in real-world photography and sharing scenarios. We therefore collect multi-image posts from RedNote, where images within the same post form a naturally co-occurring collection. To further cover other related photographs encountered in real-world photography, such as images captured under changes in illumination or viewpoint, we additionally incorporate data from HPatches (Balntas et al., 2017) and ETH3D (Schöps et al., 2017; 2019). The latter arises from image-editing processes, where multiple edited variants originate from a common source image. We therefore collect edited-image data from Lightroom to represent this form of contextual relatedness. To control the scale of individual collections, those exceeding a predefined size limit are subsampled to remain within the limit.

We formulate aesthetic annotation as pairwise preference comparisons, which directly captures relative aesthetic judgments without requiring annotators to assign absolute aesthetic scores (Ailon, 2008). Given a collection of n images, exhaustive pairing would require $\binom{n}{2}$ comparisons, making the annotation cost grow rapidly with collection size. We therefore construct a regular graph (Wong & Coppersmith, 1974) over the images in each collection and use collection-size-dependent step sizes to determine the paired edges, as illustrated in Figure 1(b). This deterministic strategy avoids random pair sampling, balances the participation of images in comparisons, and maintains efficient connectivity and broad coverage of within-collection relationships with substantially fewer pairs than exhaustive enumeration. The detailed construction procedure and parameterization are provided in the Appendix.

3.2 Preference Annotation

We recruited 15 annotators to provide pairwise aesthetic preference judgments. To distribute the annotation workload while maintaining consistency across batches, the annotators are organized into overlapping groups of 10 through a cyclic rotation scheme, such that adjacent batches share most of their annotators. To mitigate positional bias, the presentation order of each image pair is randomized during annotation. Quality-control items are interspersed throughout the annotation process to monitor annotation reliability. All annotators performed well on these control items and met the predefined quality criteria.

After annotation, pairs receiving fewer than 10 valid judgments are discarded to ensure sufficient annotation coverage. We then calibrate the pairwise preferences within each collection using the Bradley–Terry model (Bradley & Terry, 1952). The model estimates a latent aesthetic score for each image, from which we obtain the estimated score difference Δ and its standard error SE for each pair. Within each collection, we use $|\Delta|/SE$ as a measure of the confidence of each pairwise

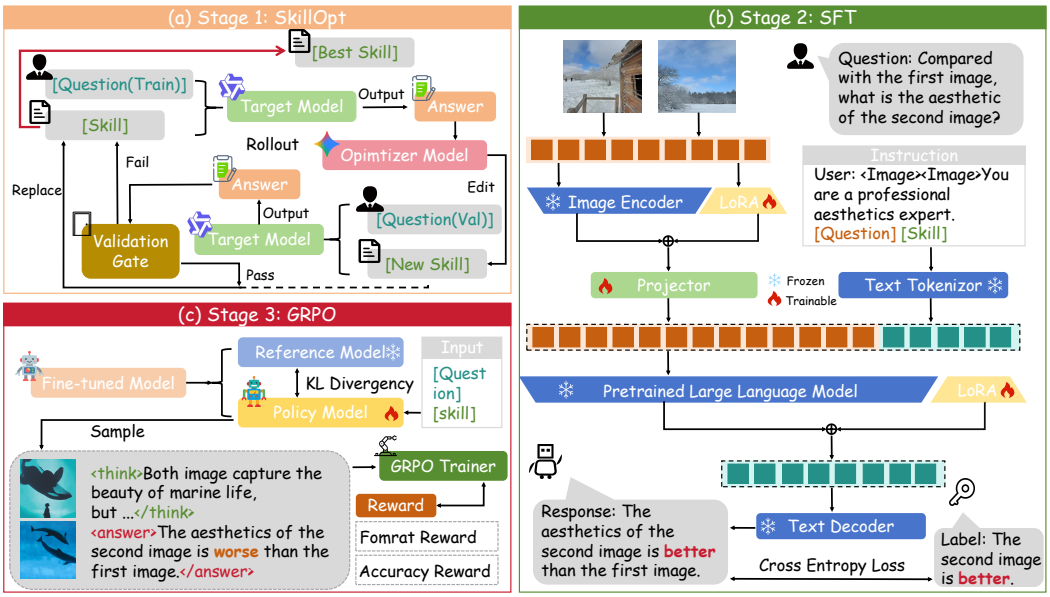


Figure 4: Three-Stage Training Pipeline for ContextAes. (a) SkillOpt provides useful prior knowledge. (b) SFT enhances aesthetic capability. (c) GRPO further aligns predictions with human aesthetic preferences.

preference and remove the pairs with the lowest values, retaining the remaining comparisons as the final preference annotations. Details are provided in the Appendix.

3.3 Dataset Analysis

To further examine aesthetic preferences in real-world image collections, we construct **ContextAes-Album**, a dataset collected from 22 independent album providers. It contains 4,140 images organized into 494 collections. For each collection, the album provider ranks the top images they consider aesthetically preferred, providing an ordering that reflects their personal aesthetic preferences. The detailed collection and ranking procedure are provided in the Appendix.

Figure 2(a) presents the distribution of contextual relatedness across ContextAes-40K, ContextAes-Album, AVA (Murray et al., 2012), and FGAesthetics (Yang et al., 2026b). ContextAes-40K closely follows ContextAes-Album but differs substantially from AVA and FGAesthetics, indicating that existing datasets mainly focus on independent images or highly similar image series, leaving broader contextually related collections less explored. Figure 2(b) characterizes low-, mid-, and high-level similarity using LPIPS (Zhang et al., 2018), DreamSim (Fu et al., 2023), and CLIP (Hessel et al., 2021), respectively. ContextAes-40K closely matches the real-world distribution of ContextAes-Album across all three measures, while AVA and FGAesthetics support this observation.

4 ContextAes

We propose **ContextAes**, an MLLM-based model for aesthetic preference assessment. As illustrated in Figure 4, it adopts a three-stage training strategy: SkillOpt (Yang et al., 2026a) provides strong prior aesthetic knowledge for subsequent training, SFT develops the model's capability for aesthetic preference assessment, and GRPO (Shao et al., 2024) further aligns the model with human preferences.

4.1 Skill Optimization

Task-specific skills can provide useful prior knowledge for large multimodal models (Panigrahi et al., 2023; Wang et al., 2023; Kuroki et al., 2025). However, conventional skills are often manually crafted or generated in a one-shot manner, limiting their quality and iterative refinement (Alzubi

et al., 2026; Ni et al., 2026). Recent work on SkillOpt addresses this issue by treating a natural-language skill as a trainable artifact associated with a frozen target model, allowing the skill to be iteratively optimized with controlled updates and validation-based selection (Yang et al., 2026a).

We adopt SkillOpt to provide strong prior aesthetic knowledge for ContextAes, as illustrated in Figure 4(a). We select Qwen3.5-9B as the target model and the backbone for subsequent training, leveraging its native multimodal architecture and strong visual understanding capability (Team, 2026). A stronger closed-source model, Gemini 3.1 Pro (Team et al., 2023), is used as the optimizer model to generate and refine the skill. Starting from an empty skill document, the optimizer iteratively updates the skill based on training performance and retains only candidates that improve performance on a held-out validation set. After convergence, the best-performing skill is selected for supervised fine-tuning.

4.2 Supervised Fine-Tuning

Supervised fine-tuning (SFT) has been widely adopted to adapt pretrained multimodal models to specific tasks and establish task-specific capabilities (Ouyang et al., 2022). We therefore employ SFT to strengthen the model's aesthetic capability following the skill optimization stage, as illustrated in Figure 4(b). With the learned aesthetic skill incorporated into the input instruction, we optimize the model using the standard autoregressive cross-entropy loss. To improve training efficiency while preserving the pretrained representations, we apply LoRA (Hu et al., 2021) to the visual and language modules, whereas the intermediate merger module is fully fine-tuned to facilitate effective fusion and alignment of visual and textual features.

4.3 Preference Optimization with GRPO

Recent reinforcement learning approaches have demonstrated strong effectiveness in eliciting reasoning capabilities and aligning model behavior with preference signals (Ouyang et al., 2022; Guo et al., 2025). GRPO (Shao et al., 2024), in particular, estimates the relative quality of multiple responses sampled for the same input and does not require an additional value or critic model, making it suitable for preference optimization with verifiable rewards. We therefore further optimize the model obtained from SFT with GRPO to better align its predictions with human aesthetic preferences, as illustrated in Figure 4(c).

To increase the diversity of preference training samples and mitigate positional bias, each annotated image pair is reversed to form an additional training instance, with its preference label reversed accordingly. This procedure doubles the number of training instances while exposing the model to both presentation orders.

For each generated response, we define the reward according to its output format and preference correctness:

$$r = \begin{cases} 0, & \text{if the output format is invalid,} \\ \lambda_0, & \text{if the format is valid but the preference is incorrect,} \\ \lambda_0 + \lambda_1, & \text{if the format is valid and the preference is correct,} \end{cases} \tag{1}$$

where λ_0 denotes the reward for valid formatting, while λ_1 denotes the additional reward for correct preference prediction.

For each training instance, the model samples a group of G responses. The rewards are normalized within the group to obtain the relative advantage:

$$A_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon}, \qquad \mu_r = \frac{1}{G}\sum_{j=1}^{G} r_j, \tag{2}$$

where σ_r is the standard deviation of the group rewards and ϵ is a small constant for numerical stability.

Table 1: Performance comparisons on ContextAes-40K under the *Natural* and *Editing* scenarios, and on the combined set (*Overall*). The best performance is **bolded**, and the second best is underlined.

Model	Natural		Editing		Overall	
	Acc↑	*F1*↑	*Acc*↑	*F1*↑	*Acc*↑	*F1*↑
Open-source MLLMs						
Qwen2.5-VL-7B	55.37%	50.87%	59.54%	54.39%	57.27%	52.47%
Qwen3-VL-8B	64.08%	63.86%	70.20%	70.19%	66.87%	66.78%
Qwen3.5-9B	65.17%	65.08%	69.22%	68.64%	67.02%	66.75%
InternVL3-8B	64.11%	63.56%	66.37%	65.83%	65.14%	64.59%
InternVL3.5-8B	61.96%	61.04%	68.16%	67.91%	64.78%	64.23%
InternVL3.5-30B	60.24%	53.22%	57.26%	49.89%	58.89%	51.70%
MiniCPM-V4.6-1B	64.33%	63.84%	63.01%	61.09%	63.74%	62.74%
LLaVA-OV2-8B	61.17%	60.54%	64.25%	62.06%	62.57%	61.35%
Kimi-VL-16B	60.82%	59.31%	65.68%	63.23%	63.03%	61.11%
GLM-4.6V-16B	60.87%	59.44%	68.83%	67.91%	64.50%	63.32%
Closed-source MLLMs						
Gemini2.5-Pro	59.52%	57.16%	63.41%	62.73%	61.34%	59.93%
Gemini3.1-Pro	64.01%	59.99%	68.83%	67.91%	66.28%	64.06%
IAA Models						
Q-Align	61.39%	61.26%	59.41%	59.41%	60.49%	60.45%
FGAes	59.40%	59.27%	57.04%	57.04%	58.32%	58.28%
ArtiMuse	56.38%	56.31%	53.10%	53.10%	54.89%	54.87%
ContextAes (Ours)	**72.44%**	**72.44%**	**78.71%**	**78.60%**	**75.30%**	**75.27%**

The resulting advantages are used to optimize the model with the GRPO objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t}\min\Big(\rho_{i,t}A_i,\ \text{clip}(\rho_{i,t}, 1-\epsilon_c, 1+\epsilon_c)A_i\Big) - \beta D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right)\Bigg]. \tag{3}$$

where θ denotes the model parameters, $|o_i|$ is the number of tokens in the i-th response, ϵ_c is the clipping coefficient, and β controls the strength of the KL regularization. Here, π_θ denotes the current policy, while π_{ref} is the reference policy.

The probability ratio is $\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|o_{i,<t},x)}{\pi_{\text{old}}(o_{i,t}|o_{i,<t},x)}$, where $o_{i,t}$ is the t-th token, $o_{i,<t}$ its preceding tokens, x the input, and π_{old} the policy used to generate the responses.

5 Experiments

5.1 Experimental Setup

We train ContextAes based on Qwen3.5-9B (Team, 2026) using the three-stage training strategy described above. SkillOpt is optimized with a batch size of 48 while keeping the model frozen. We then fine-tune the model with LoRA on eight 96G NVIDIA H20 GPUs, with a global batch size of 128, followed by GRPO optimization with a global batch size of 16. The reward coefficients are set to $\lambda_0 = 0.3$ and $\lambda_1 = 1.0$, with a generation group size of 4 and a KL coefficient of $\beta = 0.001$. Detailed optimization configurations are provided in the Appendix.

5.2 Main Results on ContextAes-40K

Evaluation Setting. We split ContextAes-40K into training and test sets with a ratio of 5:1. We evaluate 16 models in total, including 10 open-source general-purpose MLLMs: Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), Qwen3.5-9B (Team, 2026), InternVL3-8B (Zhu et al., 2025), InternVL3.5-8B and InternVL3.5-30B (Wang et al., 2025), MiniCPM-V4.6-1B (Yu et al., 2026), LLaVA-OV2-8B (An et al., 2026), Kimi-VL-16B (Team et al., 2025b), and GLM-4.6V-16B (Team et al., 2025a); 2 closed-source MLLMs: Gemini 2.5 Pro (Comanici et al., 2025) and Gemini 3.1 Pro (Team et al., 2023); 3 specialized IAA models: Q-Align (Wu et al.,

Table 2: Cross-dataset comparisons. The best and second-best performances are highlighted in **bold** and underlined, respectively. † Results are taken directly from the original papers as these models cannot be tested. * Results are trained only on AVA to compare generalization ability. For models without scoring capability, we prompt them to directly output scores as text for evaluation.

Model	AVA		PARA		TAD66K		FLICKR-AES		ArtiMuse-10K	
	SRCC	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*
Traditional Models										
MUSIQ	0.225	0.258	0.490	0.600	0.099	0.149	0.150	0.216	-0.060	-0.074
TANet†	0.758	0.765	–	–	0.513	0.531	–	–	–	–
VILA	0.776	0.775	0.651	0.658	0.418	0.444	0.616	0.645	0.273	0.268
AesMamba†	0.774	0.769	**0.936**	0.902	0.511	0.483	–	–	–	–
MLLMs for General-Purpose Applications										
mPLUG-Owl2	0.206	0.211	0.376	0.372	0.089	0.106	0.382	0.359	0.159	0.145
ShareGPT-4V	0.213	0.199	0.509	0.417	0.097	0.091	0.335	0.289	0.076	0.057
Qwen2.5-VL-7B	0.391	0.371	0.721	0.743	0.240	0.242	0.621	0.578	0.256	0.179
InternVL3-8B	0.364	0.332	0.667	0.693	0.203	0.191	0.553	0.459	0.187	0.157
IAA MLLMs										
Q-Instruct	0.318	0.338	0.569	0.724	0.122	0.159	0.259	0.299	-0.045	-0.056
PEAS	0.748	0.748	0.686	0.700	0.415	0.444	0.577	0.613	0.306	0.293
Q-Align	0.822	0.817	0.913	0.888	0.501	0.531	0.798	0.818	0.551	0.573
UNIAA-LLaVA†	0.713	0.704	0.864	0.895	0.411	0.425	0.724	0.751	–	–
Next Token Is Enough†	0.828	0.825	–	–	0.413	0.444	–	–	–	–
ArtiMuse	0.827	0.826	**0.936**	**0.958**	0.510	0.543	0.814	0.837	0.614	0.627
ContextAes (Ours)	**0.832**	**0.831**	0.933	0.956	**0.516**	**0.547**	**0.819**	**0.839**	**0.617**	**0.641**
Comparison of Generalization Ability										
Q-Align*	0.822	0.817	0.694	0.711	0.417	0.445	0.643	0.664	0.337	0.320
ArtiMuse*	0.827	0.826	0.697	0.725	0.419	0.451	0.647	0.676	0.395	0.376
ContextAes (Ours)*	0.832	0.831	0.702	0.733	0.437	0.469	0.672	0.698	0.417	0.405

2023), FGAes (Yang et al., 2026b), and ArtiMuse (Cao et al., 2026); and our ContextAes. Detailed descriptions of the compared models are provided in the Appendix. Performance is evaluated from three perspectives: *Natural*, *Editing*, and *Overall*. We report Accuracy (Acc) and F1 score as the evaluation metrics. The complete results are presented in Table 1 and Figure 3.

Results. ContextAes achieves the best performance across all three settings, improving overall Accuracy and F1 by 8.28 and 8.49 percentage points, respectively, over the strongest competing results. The gains are consistent across *Natural* and *Editing*, with a larger improvement on the latter. Specialized IAA models underperform general-purpose MLLMs, suggesting that conventional aesthetic models may not fully capture preferences in related image collections.

5.3 Cross-Dataset Evaluation

Evaluation Setting. We evaluate ContextAes on five widely used image aesthetic assessment benchmarks, including AVA (Murray et al., 2012), PARA (Yang et al., 2022), TAD66K (He et al., 2022), FLICKR-AES (Ren et al., 2017), and ArtiMuse-10K (Cao et al., 2026), using SRCC and PLCC as evaluation metrics. We compare 15 methods in total, including 4 traditional models: MUSIQ (Ke et al., 2021), TANet (He et al., 2022), VILA (Lin et al., 2024), and AesMamba (Gao et al., 2024); 4 general-purpose MLLMs: mPLUG-Owl2 (Ye et al., 2024), ShareGPT-4V (Chen et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025b), and InternVL3-8B (Zhu et al., 2025); 6 specialized IAA MLLMs: Q-Instruct (Wu et al., 2024), PEAS (Yun & Choo, 2024), Q-Align (Wu et al., 2023), UNIAA-LLaVA (Zhou et al., 2024), Next Token Is Enough (Li et al., 2025), and ArtiMuse (Cao et al., 2026); and our ContextAes. Detailed descriptions of the compared methods are provided in the Appendix. Q-Align, ArtiMuse, and ContextAes are fine-tuned on each target dataset. The complete results are presented in Table 2.

Results. ContextAes achieves the best or second-best performance across all five benchmarks. It achieves the highest SRCC and PLCC on AVA, TAD66K, FLICKR-AES, and ArtiMuse-10K, while obtaining the second-best performance on PARA, where ArtiMuse performs slightly better. These results show that ContextAes maintains strong aesthetic assessment performance across a diverse range of existing benchmarks, despite being developed for a different evaluation setting.

Table 3: Ablation study. The best performance is **bolded**.

No.	Training Strategy			Natural		Editing		Overall	
	SkillOpt	SFT	GRPO	*Acc*↑	*F1*↑	*Acc*↑	*F1*↑	*Acc*↑	*F1*↑
(1)		✓		72.00%	71.99%	76.99%	76.92%	74.27%	74.27%
(2)	✓			69.67%	69.59%	74.30%	73.60%	72.19%	71.84%
(3)			✓	69.41%	69.35%	74.20%	74.19%	71.59%	71.57%
(4)	✓	✓		71.78%	71.66%	77.28%	77.13%	74.29%	74.15%
(5)	✓		✓	67.85%	67.51%	73.74%	73.73%	70.52%	70.40%
(6)		✓	✓	69.31%	68.93%	75.46%	75.30%	72.11%	71.85%
(7)	✓	✓	✓	**72.44%**	**72.44%**	**78.71%**	**78.60%**	**75.30%**	**75.27%**

Table 4: Zero-shot performance comparisons on ContextAes-Album. ‡ denotes closed-source models and § denotes task-specific IAA models; the remaining are open-source MLLMs. The best performance is **bolded**.

Model	*Acc*↑	*F1*↑	Model	*Acc*↑	*F1*↑
Qwen2.5-VL-7B	57.58%	54.72%	Kimi-VL-16B	58.81%	57.44%
Qwen3-VL-8B	58.82%	58.73%	GLM-4.6V-16B	58.82%	58.15%
Qwen3.5-9B	58.82%	58.56%	Gemini2.5-Pro‡	58.01%	56.52%
InternVL3-8B	58.77%	58.58%	Gemini3.1-Pro‡	59.13%	58.11%
InternVL3.5-8B	58.28%	56.97%	Q-Align§	56.39%	56.39%
InternVL3.5-30B	55.87%	52.00%	FGAes§	51.69%	51.69%
MiniCPM-V4.6-1B	58.23%	58.23%	ArtiMuse§	57.34%	57.34%
LLaVA-OV2-8B	56.88%	55.79%	**ContextAes** (Ours)	**63.64%**	**63.54%**

Generalization. We evaluate cross-dataset generalization by fine-tuning Q-Align, ArtiMuse, and ContextAes only on AVA and testing on the remaining four benchmarks without further adaptation. ContextAes achieves the highest SRCC and PLCC on all four benchmarks and retains the largest proportion of its AVA performance, indicating greater robustness to distribution shifts.

5.4 Ablation Study

We conduct ablation studies to examine the contribution of each training stage, with results shown in Table 3. SFT alone provides a strong foundation, while SkillOpt alone is less effective; combining them further improves performance, indicating complementary benefits. GRPO does not consistently improve performance without SkillOpt, whereas the full three-stage strategy achieves the best performance across all settings.

5.5 Real-World Evaluation

For real-world zero-shot evaluation on ContextAes-Album, we convert user-provided rankings into pairwise preference comparisons by exhaustively pairing ranked and unranked images. Each pair is also reversed with its label swapped to balance presentation order. Using all resulting pairs as the test set, we evaluate Accuracy and F1. As shown in Table 4, ContextAes achieves the best performance among all compared methods, demonstrating its ability to capture aesthetic preferences in real-world user-curated collections.

6 Conclusion

This work explores aesthetic preference assessment within contextually related image collections. We introduce ContextAes-40K , a benchmark with pairwise preference annotations across diverse contextual relationships, and ContextAes, an MLLM-based model with a three-stage training strategy for aesthetic preference learning. We further construct ContextAes-Album as an independent real-world benchmark based on user-curated photo collections and user-provided rankings. Experiments on ContextAes-40K, established aesthetic benchmarks, and the independent real-world ContextAes-Album demonstrate strong within-collection preference prediction, cross-dataset generalization, and real-world applicability. These results highlight the importance of relative aesthetic preference learning beyond conventional IAA.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the paper, including improving grammar, clarity, fluency, and wording. We have not used generative AI tools to design or conduct experiments, generate or modify research data, or formulate research hypotheses. Other uses of generative AI for research purposes are not applicable to this work. We have reviewed all AI-assisted work and manually verified the accuracy, consistency, and appropriateness of the resulting text. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide sufficient details to facilitate the reproducibility of our work. The model architecture, training strategy, and experimental settings are described in the main paper, while additional details on dataset construction, annotation, and implementation are provided in the appendix and supplementary materials.

REFERENCES

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11569–11579, 2021.

Nir Ailon. Reconciling real scores with binary comparisons: A new logistic based model for ranking. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 21, 2008.

Salaheddin Alzubi, Noah Provenzano, Jaydon Bingham, Weiyuan Chen, and Tu Vu. Evoskill: Automated skill discovery for multi-agent systems. *arXiv preprint arXiv:2603.02766*, 2026.

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

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15313–15322, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *Proceedings of the European Conference on Computer Vision*, pp. 370–387. Springer, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Venus: Benchmarking and empowering multimodal large language models for aesthetic guidance and cropping. *arXiv preprint arXiv:2602.23980*, 2026.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683. IEEE, 2020.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Fei Gao, Yuhao Lin, Jiaqi Shi, Maoying Qiao, and Nannan Wang. Aesmamba: Universal image aesthetic assessment with state space models. In *Proceedings of the ACM International Conference on Multimedia*, pp. 7444–7453, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, volume 3, pp. 6, 2022.

Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, and Huadong Ma. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision*, pp. 21781–21790. IEEE, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the ACM International Conference on Multimedia*, pp. 5911–5920, 2024a.

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

Qifei Jia, Xintong Yao, Yasen Zhang, Minghao Li, Yajie Chai, Qiming Lu, Baoyue Shen, Runyu Shi, Ying Huang, and Yue Zhang. Beyond absolute scores: Relative edit-induced difference for generalizable image aesthetic assessment. *arXiv preprint arXiv:2606.05778*, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*, pp. 5128–5137. IEEE, 2021.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Proceedings of the European Conference on Computer Vision*, pp. 662–679. Springer, 2016.

So Kuroki, Taishi Nakamura, Takuya Akiba, and Yujin Tang. Agent skill acquisition for large language models via cycleqd. In *Proceedings of the International Conference on Learning Representations*, volume 2025, pp. 47327–47348, 2025.

Mingxing Li, Rui Wang, Lei Sun, Yancheng Bai, and Xiangxiang Chu. Next token is enough: Realistic image quality and aesthetic scoring with multimodal large language model. *arXiv preprint arXiv:2503.06141*, 2025.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26679–26689. IEEE, 2024.

Alexander C Loui and Andreas Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *IEEE Transactions on Multimedia*, 5(3):390–402, 2003.

Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *Proceedings of the 2011 International Conference on Computer Vision*, pp. 2206–2213. IEEE, 2011.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415. IEEE, 2012.

Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Trace2skill: Distill trajectory-local lessons into transferable agent skills. *arXiv preprint arXiv:2603.25158*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Proceedings of the Advances in neural information processing systems*, volume 35, pp. 27730–27744, 2022.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. In *Proceedings of the International Conference on Machine Learning*, pp. 27011–27033. PMLR, 2023.

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *2017 IEEE International Conference on Computer Vision*, pp. 638–647. IEEE, 2017.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017.

Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2019.

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.

Wei-Tse Sun, Ting-Hsuan Chao, Yin-Hsi Kuo, and Winston H Hsu. Photo filter recommendation by category-aware aesthetic learning. *IEEE Transactions on Multimedia*, 19(8):1870–1880, 2017.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian,

Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025a. URL https://arxiv.org/abs/2508.06471.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

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

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

CK Wong and Don Coppersmith. A combinatorial problem related to multimodule memory organizations. *Journal of the ACM*, 21(3):392–402, 1974.

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. *arXiv preprint arXiv:2312.17090*, 2023.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25490–25500. IEEE, 2024.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, et al. Skillopt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026a.

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19829–19837. IEEE, 2022.

Zhichao Yang, Jianjie Wang, Zhixianhe Zhang, Pangu Xie, Xiangfei Sheng, Pengfei Chen, and Leida Li. Fine-grained image aesthetic assessment: Learning discriminative scores from relative ranks. *arXiv preprint arXiv:2603.03907*, 2026b.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. Mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13040–13051. IEEE, 2024.

Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22388–22397. IEEE, 2023.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Jooyeol Yun and Jaegul Choo. Scaling up personalized image aesthetic assessment via task vector customization. In *European Conference on Computer Vision*, pp. 323–339. Springer, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595. IEEE, 2018.

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 *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14411–14421. IEEE, 2025.

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

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.

APPENDIX

A EXTENDED RELATED WORK

As summarized in Table 5, image aesthetic assessment (IAA) datasets have evolved substantially in terms of both assessment unit and annotation design. Early datasets predominantly focused on individual images and relied on holistic aesthetic scores, establishing the standard paradigm for image-level aesthetic assessment. Subsequent efforts extended this paradigm by enriching the annotations associated with individual images, such as fine-grained aesthetic attributes and textual descriptions, while also exploring more specialized aspects of aesthetics, including semantic themes and image color aesthetics, as exemplified by TAD66K (He et al., 2022) and ICAA17K (He et al., 2023). More recent work has begun to consider aesthetic relationships among multiple images. FGAesthetics (Yang et al., 2026b), for example, introduces aesthetic rankings within highly similar image series, moving beyond independent image assessment toward comparative evaluation. However, such relational assessment remains largely focused on tightly constrained image similarities. In contrast, our ContextAes-40K considers collections of contextually related images with broader inter-image relationships, and provides pairwise aesthetic preference annotations within each collection. This extends aesthetic assessment from independent images and highly similar series toward contextually related image collections.

B CONTEXTAES-40K: CONSTRUCTION AND ANNOTATION

To provide further details on the construction of ContextAes-40K, we describe the complete dataset development pipeline, including data collection, pairwise preference annotation, and subsequent data processing. We detail how contextually related image collections are gathered and organized, how pairwise aesthetic preferences are collected from human annotators, and how the resulting annotations are modeled and filtered to obtain the final training data.

B.1 DATA COLLECTION

To construct ContextAes-40K, we consider two complementary types of contextual relationships: naturally co-occurring relationships in real-world photography and editing-induced relationships

Table 5: Overview of representative image aesthetic assessment datasets.

Dataset	Year	# Images	Image Source	Assessment Unit	Annotation
CUHK-PQ (Luo et al., 2011)	2011	17,673	Natural	Single	Classification
AVA (Murray et al., 2012)	2012	255,530	Natural	Single	MOS
AADB (Kong et al., 2016)	2016	10,000	Natural	Single	MOS + Attributes
SPAQ (Fang et al., 2020)	2020	111,125	Natural	Single	MOS
ArtEmis (Achlioptas et al., 2021)	2021	80,031	Art	Single	MOS + Description
TAD66K (He et al., 2022)	2022	66,327	Natural/Art	Single	MOS
PARA (Yang et al., 2022)	2022	31,220	Natural	Single	MOS
BAID (Yi et al., 2023)	2023	60,337	Art	Single	MOS
ICAA17K (He et al., 2023)	2023	17,726	Natural	Single	MOS
AesBench (Huang et al., 2024b)	2024	2,800	Natural/AI/Art	Single	Description
AesGuide (Du et al., 2026)	2024	10,748	Natural/Art/AI	Single	MOS + Guidance
ArtiMuse-10K (Cao et al., 2026)	2024	10,000	Natural/Art/Design/AI	Single	MOS + Attributes + Description
FGAesthetics (Yang et al., 2026b)	2025	32,217	Natural/AI/Crop	Series	Ranking
ContextAes-40K (Ours)	**2026**	**39,033**	**Natural/Editing**	**Collection**	**Preference**
ContextAes-Album (Ours)	**2026**	**4,140**	**User-curated Photos**	**Collection**	**User Ranking**

arising from a common source image. We collect data from four sources, including RedNote, HPatches (Balntas et al., 2017), ETH3D (Schöps et al., 2017; 2019), and Lightroom. The details of each source are described below.

RedNote. RedNote is a lifestyle-oriented social media platform where users share daily experiences, interests, and visual content. Its posts commonly document a particular activity, place, event, or personal experience, providing a natural organization of multiple photographs within a shared context. We collect multi-image posts from RedNote and treat the images associated with each post as a naturally co-occurring collection. Such collections provide diverse real-world photographs whose relationships arise from the way they are captured and presented rather than from artificially constructed similarity.

ETH3D. ETH3D is a benchmark for multi-view stereo and 3D reconstruction, containing high-resolution DSLR images and multi-camera sequences captured across a variety of indoor and outdoor scenes. Its reconstruction-oriented design requires multiple views of the same physical scene, resulting in groups of photographs that depict a common environment from different camera positions. We treat the images associated with each scene as a collection, providing naturally related photographs with variations in viewpoint, visibility, and scene composition.

HPatches. HPatches is a benchmark for evaluating local image descriptors and image matching, built from image sequences exhibiting controlled viewpoint or illumination changes. Each sequence contains one reference image and five target images of the same scene, with 59 sequences involving viewpoint changes and 57 involving illumination changes. We use the images within each sequence as a collection, introducing contextual relationships arising from changes in viewpoint and illumination while preserving the underlying scene content.

Lightroom. Lightroom provides an online community where photographers share their photographs together with their editing processes. Community posts can expose the editing settings applied to an image, and the platform also supports Remix, allowing other users to create alternative edited versions of a shared photograph. We crawl photographs and their associated edited versions from Lightroom, where multiple versions derived from the same source image form an editing-induced collection. These collections provide aesthetically related images that share the same underlying visual content while exhibiting differences introduced by different editing operations.

Collectively, these sources yield a large number of image collections with varying sizes. We first remove collections containing only a single image, as such collections do not provide inter-image relationships and therefore cannot support aesthetic preference assessment. For collections containing more than 18 images, we further reduce their size through uniform sampling along the original image sequence. Specifically, let a collection contain N images $(N > 18)$ in its original order. Using zero-based indexing, the indices of the 18 retained images are determined by

$$i_j = \left\lfloor j \times \frac{N}{18} \right\rfloor, \quad j \in 0, 1, 2, \ldots, 17. \tag{4}$$

This strategy ensures that the retained images are distributed as evenly as possible across the original collection while preserving its overall ordering. Restricting collections to at most 18 images not

only provides a unified collection size for subsequent processing, but also offers a suitable scale for constructing efficient and well-connected comparison graphs during pairwise annotation.

With the collection sizes controlled, we next construct pairwise comparisons for aesthetic preference annotation. We formulate aesthetic annotation as pairwise preference comparisons, which directly capture relative aesthetic judgments without requiring annotators to assign absolute aesthetic scores (Ailon, 2008). Given a collection of N images, exhaustive pairing would require $\binom{N}{2}$ comparisons, making the annotation cost grow quadratically with collection size. To obtain a compact yet well-connected set of comparisons, we model each image as a node and each selected image pair as an undirected edge, and construct a regular graph over the images within each collection. An edge between two nodes indicates that the corresponding image pair is presented to annotators for aesthetic preference comparison.

Following the graph construction strategy Wong & Coppersmith (1974), we consider a regular graph with k independent connection step sizes. Let N denote the number of image nodes in a collection, k the number of independent step sizes, M the graph diameter, and A the average shortest-path length. Their analysis provides theoretical lower bounds and asymptotic estimates for the diameter and average path length of such networks:

$$M \geq \frac{1}{2}(k!N)^{\frac{1}{k}} - \frac{1}{2}(k+1) \approx \frac{1}{2}(k!)^{\frac{1}{k}} N^{\frac{1}{k}}, \tag{5}$$

$$A \geq \frac{k}{2(k+1)!N}\left[(k!N)^{\frac{1}{k}} - (3k+1)\right]^{k+1} \approx \frac{1}{2}\left(\frac{k}{k+1}\right)(k!)^{\frac{1}{k}} N^{\frac{1}{k}}. \tag{6}$$

For the case of $k = 2$, corresponding to two independent step sizes, the analysis indicates that the second step should be chosen on the order of $\sqrt{N}$ to obtain a highly efficient network topology (Wong & Coppersmith, 1974). We therefore use two step sizes for each collection: a basic adjacent connection with step size 1 and an additional long-range connection with step size

$$s = \left\lfloor \sqrt{N} \right\rceil, \tag{7}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer. The resulting graph combines local neighboring comparisons with long-range comparisons, providing broad coverage of within-collection relationships while reducing the number of comparisons from $O(N^2)$ to $O(N)$.

With the two-step regular graph defined above, $N \leq 18$ and $s = \lfloor \sqrt{N} \rceil$ ensure a graph diameter no larger than **3** and an average shortest-path length close to **2**. These properties ensure that images within each collection remain closely connected through a small number of comparison steps, while effectively preserving the contextual structure and inter-image relationships of the original collection.

B.2 Pairwise Preference Annotation

We recruited 15 annotators (10 male and 5 female) to provide pairwise aesthetic preference judgments. We randomly divided the collected image collections into batches, and 10 annotators were assigned to each batch. To maintain consistency across batches and mitigate potential annotation drift, the annotator assignments were designed to overlap across adjacent batches, so that each batch shared a substantial number of annotators with neighboring batches.

During annotation, the presentation order was randomized separately for each annotator using the annotator's name as the random seed. As a result, different annotators were presented with different pair orders and different left–right arrangements of the two images, which helps mitigate positional bias and systematic effects caused by prolonged annotation. In addition, attention-check items were interspersed throughout the annotation process. These items consisted of pairs with clear aesthetic quality differences, such as a high-quality image paired with a severely distorted image, and were used to identify inattentive annotation. The error rate of every annotator on the attention-check items remained below 10% across all batches, indicating consistent and reliable annotation behavior. An example of the annotation interface is shown in Figure 5.

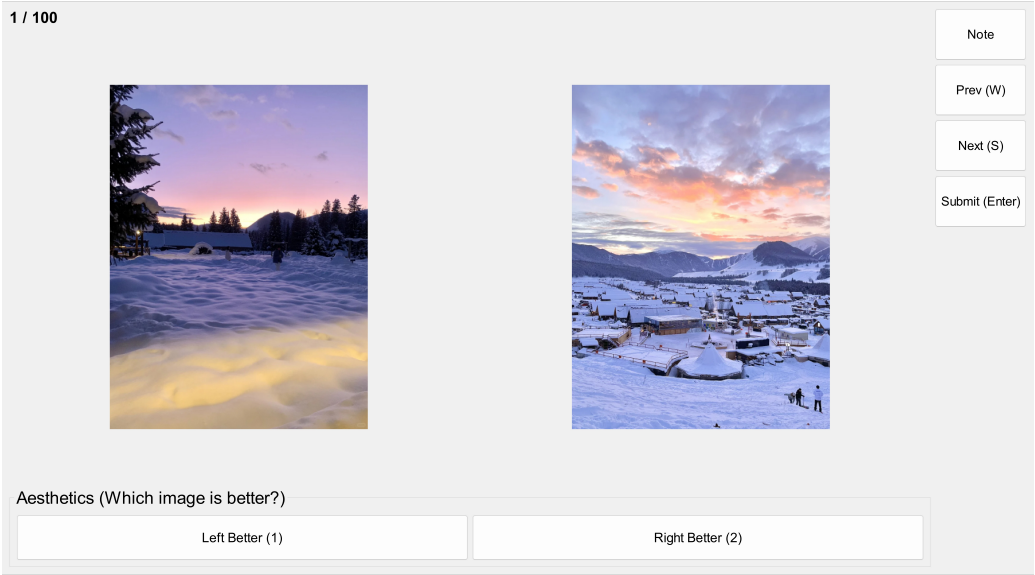


Figure 5: Annotation interface for pairwise aesthetic preference annotation.

B.3 Data Processing

After annotation, each image pair is associated with multiple pairwise aesthetic preference judgments. To ensure sufficient annotation coverage, we first discard pairs that receive fewer than 10 valid judgments. For the remaining pairs, we further calibrate the pairwise preferences within each collection using the Bradley–Terry (BT) model (Bradley & Terry, 1952).

Given a collection containing n images, the BT model assigns a latent aesthetic score s_i to each image i. For an image pair (i, j), the probability that image i is preferred to image j is modeled as

$$P(i \succ j) = \frac{\exp(s_i)}{\exp(s_i) + \exp(s_j)} = \sigma(s_i - s_j), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Let w_{ij} and w_{ji} denote the numbers of valid judgments in which image i and image j are preferred, respectively. The BT model estimates the latent scores by maximizing the corresponding log-likelihood:

$$\mathcal{L}(\mathbf{s}) = \sum_{i \neq j} w_{ij} \log P(i \succ j). \tag{9}$$

Since only score differences are identifiable, the scores are defined up to an additive constant. In practice, we impose a normalization constraint when estimating the scores.

The BT model converts the pairwise preferences within each collection into a set of calibrated latent aesthetic scores $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_n)$. The estimation procedure also provides the covariance matrix of the estimated scores, which characterizes their uncertainty. Under the standard asymptotic properties of maximum-likelihood estimation, the estimated score vector can be approximated by a multivariate normal distribution:

$$\hat{\mathbf{s}} \approx \mathcal{N}(\mathbf{s}, \boldsymbol{\Sigma}), \tag{10}$$

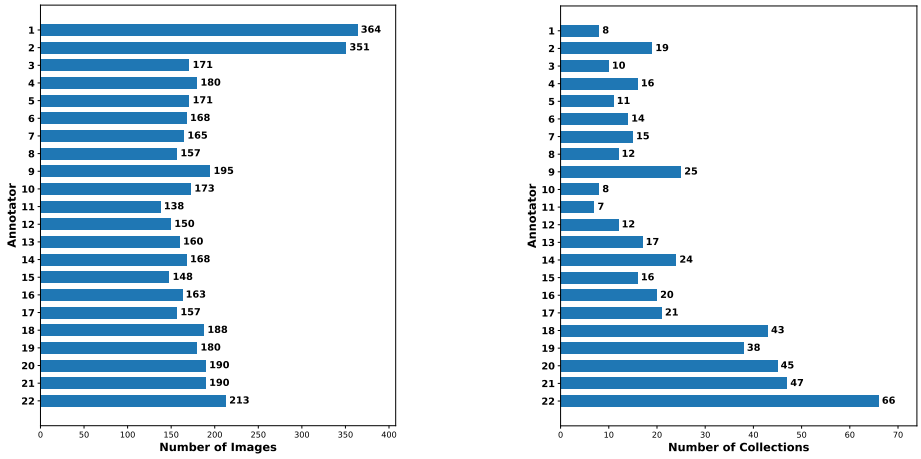
where $\boldsymbol{\Sigma}$ denotes the covariance matrix of the estimated scores.

For each pair (i, j), we calculate the estimated score difference

$$\Delta_{ij} = \hat{s}_i - \hat{s}_j. \tag{11}$$

Since Δ_{ij} is a linear combination of the estimated scores, its variance can be obtained from the covariance matrix:

$$\mathrm{Var}(\Delta_{ij}) = \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}, \tag{12}$$



(a) (b)

Figure 6: Contribution statistics of ContextAes-Album. (a) Images and (b) collections contributed by each album provider.

and the corresponding standard error is

$$SE_{ij} = \sqrt{\Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}}. \tag{13}$$

We then use the standardized score difference

$$r_{ij} = \frac{|\Delta_{ij}|}{SE_{ij}} \tag{14}$$

as a measure of the confidence of each pairwise preference. A small value of r_{ij} can arise for two reasons. First, a small $|\Delta_{ij}|$ indicates that the two images have similar estimated aesthetic scores and are therefore difficult to distinguish based on the collected preferences. Such comparisons provide relatively weak evidence of a clear preference and may introduce noise into the final annotations. Second, a large SE_{ij} indicates high estimation uncertainty, meaning that the relative preference between the two images is less precisely determined by the available judgments. Thus, a low value of $|\Delta_{ij}|/SE_{ij}$ identifies pairwise comparisons with either weak preference separation or high estimation uncertainty.

To remove such ambiguous comparisons, we perform filtering independently within each collection. Specifically, pairs are sorted according to $|\Delta_{ij}|/SE_{ij}$, and the lowest 30% are discarded. The remaining comparisons are retained as the final pairwise preference annotations of ContextAes-40K.

C ContextAes-Album: Construction and Annotation

To provide an independent external evaluation setting for real-world aesthetic preferences, we collect **ContextAes-Album**, a dataset consisting of user-curated photo collections from 22 independent album providers. Each provider contributes their own image collections, resulting in 4,140 images organized into 494 collections. The contribution statistics for each provider are shown in Figure 6.

For each contributed collection, the provider identifies the images they consider aesthetically preferred and ranks up to 5 selected images in order of preference. These rankings provide collection-level aesthetic preferences from real-world album creators. Figure 6(a) and (b) show the numbers of contributed images and collections for each provider, respectively.

D ContextAes: Model and Implementation Details

This section provides additional details of the three-stage training pipeline used to develop ContextAes, including the training procedure and implementation configurations. The pipeline consists of

SkillOpt (Yang et al., 2026a), supervised fine-tuning (SFT), and Group Relative Policy Optimization (GRPO) (Shao et al., 2024). In the first stage, SkillOpt optimizes a task-specific aesthetic skill for the target model. The resulting skill is incorporated into SFT in the second stage, followed by GRPO-based preference optimization in the final stage. The following subsections describe each stage and its corresponding training configuration in detail.

D.1 SKILL OPTIMIZATION

We adopt SkillOpt (Yang et al., 2026a) to optimize a task-specific aesthetic skill for the target model. Starting from an empty skill document, the target model performs the training tasks under the current skill, and the optimizer model then analyzes the task prompts, model responses, and corresponding evaluation feedback to propose updates to the skill. Instead of regenerating the entire skill, each update is implemented through a bounded number of add, delete, or replace operations on the current skill document. The number of allowed edits is controlled by the textual learning-rate budget, which determines the magnitude of each skill update.

In our implementation, Qwen3.5-9B is used as the target model and Gemini 3.1 Pro as the optimizer model. At each iteration, candidate skill updates are evaluated on a held-out validation set, and an updated skill is retained only when it improves the validation performance over the current skill. Starting from the empty document, this iterative optimization process continues until convergence, after which the best-performing skill on the validation set is selected for the subsequent SFT stage. The resulting best skill is given below:

 # Image Aesthetics Comparison Guide

When comparing the aesthetics of two images, prioritize natural quality, clarity, and thoughtful composition over heavy stylization. Avoid a strict bias toward "moody," "dark," or overly conventional photography.

1. Lighting and Exposure
- ***Reward** bright, clear, and inviting scenes, natural daylight, and well−lit, flattering portraits (including direct flash photography). Strongly reward dynamic, atmospheric, and cinematic lighting (e.g., golden hour, pleasing bokeh, backlit subjects with rim light, dramatic contrast, intentional silhouettes, sun flares) when it creates a strong mood. For night photography, reward cinematic lighting, intentional motion blur, and authentic urban color casts (e.g., yellow/green streetlights).*
- ***Penalize** artificially dark daytime edits, heavy vignetting, crushed shadows, and murky or grainy low−light shots. Penalize flat, dull lighting lacking artistic intent. Do not automatically assume dark, "moody ," or high−contrast lighting is superior to a bright, natural, or softly lit scene. Do not penalize beautiful golden−hour backlighting or rim light as "blown−out" if it enhances the artistic atmosphere.*

2. Color and Processing
- ***Reward** clean, accurate color representation, bright and airy palettes, and vibrant, lively colors (e.g., intense sunsets, deep blue skies). Strongly reward intentional color grading (e.g., moody cyan/green tints , warm vintage tones) and highly vibrant, energetic stylization (e.g., anime−style skies, neon) that establishes a cohesive atmosphere. Do not automatically penalize highly saturated sunsets or heavily tinted moody scenes as "unnatural" if they are visually striking.*
- ***Penalize** destructive over−processing: harsh clashing colors, extreme over−sharpening that creates halos, disjointed skin tints, heavy unnatural color shifts on wildlife/nature, artificial sun flares that wash out skies, muddy textures, and heavy, unnatural dark filters. Distinguish between a beautifully vibrant landscape and a destructively "deep−fried" edit.*
- ***Stylization vs. Natural Color:** When comparing a heavily processed, "crunchy" high−contrast, or grainy image to a softer, cleaner, more natural version of the same scene, prefer the cleaner, softer version. Do not mistake heavy, unnatural contrast for "richness." Do not penalize vibrant filters, bright flash photography, or trendy color grading as "washed out" or "over−processed."*

3. Composition and Framing
- ***Creative Perspectives & Posing:** Reward dynamic, artistic compositions, strong compositional elements (e.g., water reflections that add symmetry), beautiful backlighting, and atmospheric foreground elements. Do not penalize natural foreground elements (e.g., leaves, twigs) or wind−blown hair as " clutter" if they add depth. Prefer evocative, artistic portraits over flat, overly posed, or generic "stock− photo" style portraits. Reward tighter, well−framed crops that focus on the subject over unnecessarily wide, loose framing. Penalize highly distorted lenses, gimmicky or oversized foreground framing (e.g., a massive shoe or hands blocking the edges that dominate the frame), and genuinely distracting visual clutter.*

*– **Trendy & Candid Lifestyle:** Strongly reward bright, authentic, candid lifestyle photography, trendy social media aesthetics (e.g., mirror selfies, casual flash photography, urban backdrops), and clear, flattering portraits. Do not penalize these as "cluttered" or "washed out" if they capture a genuine moment.*
*– **Human Element & Narrative:** Strongly reward the inclusion of human subjects or figures in landscapes and environmental shots, as they add scale, narrative, and a clear focal point. Do not penalize them as "clutter" compared to an empty scene.*
*– **Authenticity & Cohesion:** Reward genuine, cohesive photographs (e.g., candid moments, visible emotional engagement) over clunky digital composites, jarring juxtapositions, or overly staged photos.*
*– **Portraits vs. Products:** For portraits, environmental shots with scenic context and natural framing are excellent, but do not penalize a tighter, well−framed crop. For ∗product photography∗, a well− composed, focused shot on a complementary background is often superior; do not automatically prefer a "lifestyle" shot if hands or skin textures introduce clutter.*
*– **Clean Geometry & Subjects:** For architectural, geometric, or macro/detail subjects, strongly reward crisp, vibrant, high−fidelity presentations. Do not automatically penalize a high−quality, sharp detail shot in favor of a moody but murky environmental shot. For artwork or flat subjects, reward straight−on, tightly framed shots that maximize detail.*

4. Clarity and Sharpness
*– **Reward** crisp, sharp, clean focus, high fidelity, and clean textures (e.g., clear water reflections, crisp night lights) that resolve fine details without being obscured by haze or noise. When comparing similar images, prefer the one with sharper details and clearer textures over a softer, hazier, or slightly blurred version.*
*– **Penalize** artificial blur, severe motion blur, excessive grain, unintentional noise, muddy textures used to fake a "vintage" or "cinematic" vibe, and heavy artificial vignetting that obscures details.*

5. Concise Reasoning & Output Format
*– **Strict Directional Labeling:** Always anchor your final decision to the prompt's question: ∗Compared with Image A, how are the aesthetics of Image B?∗*
– Conclude with a definitive choice based on the prompt's direction.
*– **CRITICAL FORMAT:** Your final output MUST be exactly one word enclosed in perfectly formed tags: either '<answer>better</answer>' or '<answer>worse</answer>' on a new line at the very end of your response. Do NOT output the word without tags, and absolutely NO extra text, summaries, or punctuation should appear outside or after these tags.*

D.2 Training Configuration

We summarize the training configurations for the three stages of ContextAes in Table 6. Unless otherwise specified, we use the default settings of the corresponding training framework.

E Additional Analysis

This section provides additional details on the models evaluated in our experiments and presents further analyses beyond those reported in the main paper. We first describe the evaluated models and then provide additional analyses of model performance across different experimental settings.

E.1 Evaluation on ContextAes-40K

E.1.1 Compared Models

We evaluate 16 models spanning three categories: open-source general-purpose MLLMs, closed-source MLLMs, and task-specific image aesthetic assessment (IAA) models. Detailed descriptions of the evaluated models are provided below.

Open-source General-purpose MLLMs.

Qwen2.5-VL-7B (Bai et al., 2025b). Qwen2.5-VL is a general-purpose vision-language model in the Qwen family. Its vision encoder uses a native dynamic-resolution ViT with window attention to reduce the computational cost of processing high-resolution images, while multimodal rotary position embedding (mRoPE) is used to model spatial and temporal information. The model supports a broad range of multimodal understanding tasks, including document analysis, visual localization, and long-video comprehension.

Table 6: Training configurations of ContextAes across the three training stages.

Skill Optimization			
Target model	Qwen3.5-9B	Optimizer model	Gemini 3.1 Pro
Batch size	48	Maximum analyst rounds	3
Edit budget	4	Minimum edit budget	2
Learning-rate scheduler	Cosine	Skill update mode	Patch-based
Supervised Fine-Tuning			
Training epochs	3	Global batch size	128
Per-device batch size	4	Gradient accumulation steps	4
Language learning rate	1×10^{-4}	Vision learning rate	2×10^{-5}
Merger learning rate	1×10^{-4}	Weight decay	0.01
Warmup ratio	0.05	LR scheduler	Cosine
LoRA rank	64	LoRA α	128
LoRA dropout	0.05	Minimum image pixels	256×28^2
Maximum image pixels	1280×28^2		
Group Relative Policy Optimization			
Training epochs	2	Per-device batch size	2
Gradient accumulation steps	16	Group size G	4
Language learning rate	2×10^{-6}	Vision learning rate	5×10^{-7}
Merger learning rate	2×10^{-6}	Weight decay	0.01
Warmup ratio	0.05	LR scheduler	Cosine
Maximum prompt length	4096	Maximum completion length	512
Minimum image pixels	256×28^2	Maximum image pixels	1280×28^2
Temperature	0.95	Top-p	0.9
Top-k	100	Repetition penalty	1.0
Reward weights (λ_0, λ_1)	$(0.3, 1.0)$	KL coefficient β	0.001

Qwen3-VL-8B (Bai et al., 2025a). Qwen3-VL is a general-purpose vision-language model designed with enhanced visual perception and multimodal reasoning capabilities. It adopts a vision encoder–MLP merger–LLM architecture and introduces DeepStack to inject multi-level visual features into the language model. Interleaved-MRoPE further supports the modeling of spatial and temporal information across multimodal inputs.

Qwen3.5-9B (Team, 2026). Qwen3.5-9B is a multimodal foundation model in the Qwen family with a unified vision-language architecture. It adopts an early-fusion design that allows visual and textual tokens to be processed within the same language-modeling framework, together with a hybrid language architecture for efficient long-context processing. It serves as the base model for ContextAes before task-specific training.

InternVL3-8B (Zhu et al., 2025). InternVL3 follows the ViT–MLP–LLM paradigm and introduces native multimodal pre-training, jointly optimizing language and multimodal capabilities during pre-training. It adopts dynamic-resolution visual processing and Variable Visual Position Encoding (V2PE) to improve multimodal context modeling. The model supports image, multi-image, video, and document understanding together with general multimodal reasoning.

InternVL3.5-8B (Wang et al., 2025). InternVL3.5 extends the InternVL architecture with further improvements in multimodal perception, reasoning, and efficiency. It retains the ViT–MLP–LLM design while employing an enhanced training pipeline that combines multimodal pre-training, supervised fine-tuning, and cascade reinforcement learning. The model is designed for diverse multimodal understanding and reasoning tasks.

InternVL3.5-30B (Wang et al., 2025). InternVL3.5-30B shares the general architecture and training paradigm of InternVL3.5 while providing substantially greater language-model capacity than the 8B variant. Its multimodal architecture combines dynamic high-resolution visual processing with a large language backbone to support more demanding visual-language reasoning tasks.

MiniCPM-V4.6-1B (Yu et al., 2026). MiniCPM-V4.6 is a lightweight multimodal model built with a SigLIP2-400M vision encoder and a Qwen3.5-0.8B language model. It introduces visual-token compression within the vision encoder and supports mixed $4\times$ and $16\times$ compression rates to

balance visual fidelity and computational efficiency. The resulting compact architecture is designed for efficient multimodal processing and deployment.

LLaVA-OV2-8B (An et al., 2026). LLaVA-OneVision-2 is a general-purpose multimodal model that combines a vision encoder, a visual-language projection module, and a language model for unified multimodal processing. Its architecture is designed to support single-image, multi-image, and video inputs within a common framework, enabling visual understanding and multimodal reasoning across different input configurations.

Kimi-VL-16B (Team et al., 2025b). Kimi-VL adopts a native-resolution visual encoder, MoonViT, together with an MLP projector and a mixture-of-experts (MoE) language model. The model contains 16B total parameters while activating only a small subset of language-model parameters for each token, providing an efficient sparse architecture. It supports long-context multimodal understanding across images, videos, and documents.

GLM-4.6V-16B (Team et al., 2025a). GLM-4.6V is a general-purpose multimodal model in the GLM-V family, designed for long-context visual understanding and multimodal reasoning. It introduces native multimodal function calling, allowing visual inputs such as images, screenshots, and document pages to participate directly in tool-oriented reasoning and execution.

Closed-source MLLMs.

Gemini 2.5 Pro (Comanici et al., 2025). Gemini 2.5 Pro is a proprietary multimodal reasoning model developed by Google. Its internal architecture is not publicly disclosed, while the model is designed to process text, image, audio, and video inputs within a unified multimodal reasoning framework. It supports long-context processing and complex multimodal reasoning.

Gemini 3.1 Pro (Team et al., 2023). Gemini 3.1 Pro is a later-generation native multimodal reasoning model in the Gemini family. It supports heterogeneous multimodal inputs together with long-context reasoning and complex task solving. It is also used as the optimizer model in the SkillOpt stage of ContextAes.

Task-specific IAA Models.

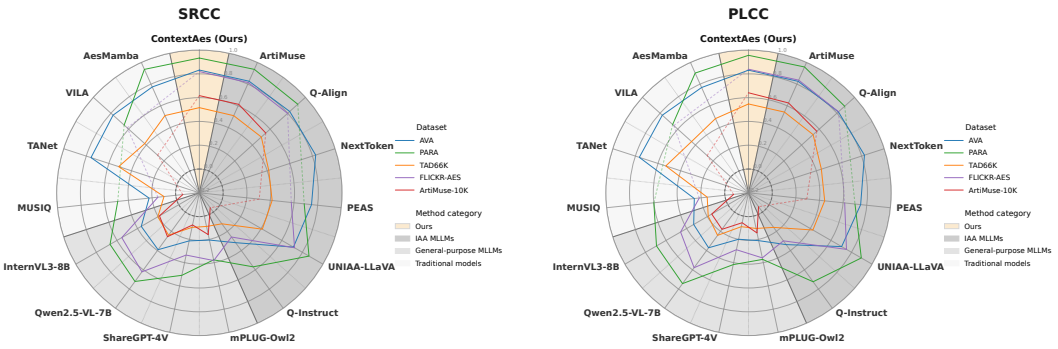
Q-Align (Wu et al., 2023). Q-Align is an LMM-based method for visual quality and aesthetic assessment that reformulates subjective scoring as classification over discrete text-defined levels. Instead of directly predicting a continuous score, the model is trained to associate visual quality with predefined textual rating levels, allowing the underlying multimodal model to leverage its language understanding capability for perceptual assessment.

FGAes (Yang et al., 2026b). FGAesQ is a fine-grained IAA model designed to capture subtle aesthetic differences between visually related images. It introduces Difference-preserved Tokenization (DiffToken), Comparative Text-assisted Alignment (CTAlign), and Rank-aware Regression (RankReg) to preserve discriminative visual differences, align them with comparative textual descriptions, and incorporate relative aesthetic ranks into score prediction. Its associated FGAesthetics dataset is constructed from image series with pairwise comparisons, making the formulation closely related to preference assessment among related images.

ArtiMuse (Cao et al., 2026). ArtiMuse is an MLLM-based IAA model that combines holistic aesthetic scoring with fine-grained aesthetic understanding. It is built on a multimodal vision-language architecture and incorporates task-specific supervision for both quantitative aesthetic assessment and detailed aesthetic analysis. Its associated ArtiMuse-10K dataset provides expert annotations over multiple aesthetic attributes together with holistic aesthetic scores.

E.1.2 Results and Analysis

Beyond the overall performance gap, several patterns can be observed from the detailed results. First, performance varies considerably across general-purpose MLLMs despite their broad multimodal capabilities. The Qwen3 family achieves the strongest results among the open-source models, with Qwen3-VL-8B reaching 66.87% Accuracy and 66.78% F1 on the Overall setting, while Qwen3.5-9B obtains the highest Overall Accuracy among the general-purpose models at 67.02%. In contrast, increasing model scale within the same family does not consistently improve performance. For example, InternVL3.5-30B performs substantially worse than InternVL3.5-8B in both scenarios,



(a) (b)

Figure 7: Cross-dataset aesthetic assessment performance. (a) SRCC and (b) PLCC.

with Overall Accuracy decreasing from 64.78% to 58.89%. This suggests that model capacity alone is insufficient for reliable aesthetic preference assessment, and that the formulation and training of the task also play an important role.

Second, the performance differences between the *Natural* and *Editing* settings reveal that the two scenarios impose different levels of difficulty on existing models. Many general-purpose MLLMs achieve higher scores on *Editing* than on *Natural*, whereas the opposite pattern is observed for several smaller or specialized models. For instance, Qwen3-VL-8B improves from 64.08% to 70.20% in Accuracy when moving from *Natural* to *Editing*, while Q-Align decreases from 61.39% to 59.41%. ContextAes maintains strong performance in both settings, reaching 72.44% Accuracy on *Natural* and 78.71% on *Editing*. The relatively consistent gains across the two scenarios indicate that the proposed training strategy is not restricted to a particular type of contextual relationship.

Third, the comparison between open-source and closed-source MLLMs shows that stronger general multimodal capability does not automatically translate into better performance on contextual aesthetic preference assessment. Gemini 3.1 Pro reaches 66.28% Overall Accuracy and 64.06% F1, which remains below the strongest open-source models despite its strong general-purpose multimodal reasoning capability. This gap suggests that broad multimodal reasoning ability and task-specific preference modeling capture different aspects of the problem. In particular, recognizing which image is preferred within a collection requires modeling relative aesthetic differences between the given images rather than evaluating each image in isolation. The relationship between Accuracy and F1 also provides an informative view of model behavior. For ContextAes, the two metrics are nearly identical across all three settings, with only a 0.17 percentage-point difference on the Overall set. A similar consistency is observed for most task-specific IAA models, whereas several general-purpose MLLMs exhibit larger gaps between Accuracy and F1, such as Qwen2.5-VL-7B and InternVL3.5-30B. This indicates that the performance gains of ContextAes are not concentrated on a particular class and remain stable across the two complementary evaluation metrics.

Finally, the comparison between Qwen3.5-9B and ContextAes provides a controlled view of the effect of task-specific training. Both models share the same underlying backbone, yet ContextAes substantially improves over Qwen3.5-9B across all settings and metrics. Since the backbone architecture remains unchanged, the difference provides direct evidence that adapting a general-purpose multimodal model specifically to contextual aesthetic preferences can produce substantial gains beyond those obtained from the pretrained model alone.

E.2 Cross-Dataset Generalization

E.2.1 Compared Methods and Evaluation Datasets

We evaluate ContextAes on five widely used image aesthetic assessment benchmarks, including AVA (Murray et al., 2012), PARA (Yang et al., 2022), TAD66K (He et al., 2022), FLICKR-AES (Ren et al., 2017), and ArtiMuse-10K (Cao et al., 2026). These benchmarks cover different aesthetic annotation schemes and image distributions, enabling us to examine the aesthetic assessment capability of the evaluated methods under varying data distributions. We compare 15 methods

in total, including traditional IAA models, general-purpose MLLMs, and specialized IAA MLLMs. Detailed descriptions of the compared methods are provided below.

Traditional IAA Models.

MUSIQ (Ke et al., 2021). MUSIQ is a multi-scale image quality Transformer designed to process images at native resolutions with varying sizes and aspect ratios. Instead of resizing all images to a fixed input shape, it constructs a multi-scale representation from the original image and additional resized versions. It further introduces hash-based two-dimensional spatial embeddings and scale embeddings to preserve spatial information across resolutions, enabling the model to capture visual quality at different granularities.

TANet (He et al., 2022). TANet, or Theme and Aesthetics Network, is a theme-aware aesthetic assessment model introduced with TAD66K. It explicitly models the relationship between image aesthetics and semantic themes and adaptively learns aesthetic prediction rules according to the recognized theme. It further incorporates an RGB-distribution-aware attention module to capture theme-dependent color characteristics while alleviating the computational burden of conventional attention mechanisms.

VILA (Lin et al., 2024). VILA learns image aesthetics from user comments through vision-language pre-training. It employs an image-text encoder-decoder architecture and jointly uses contrastive and generative objectives to learn rich aesthetic semantics from image-comment pairs without relying exclusively on numerical aesthetic labels. For downstream IAA, VILA introduces a lightweight rank-based adapter that uses text as an anchor for learning aesthetic ranking concepts.

AesMamba (Gao et al., 2024). AesMamba is a universal IAA framework based on state-space models. It replaces conventional CNN- or ViT-based visual backbones with VMamba to efficiently capture both global and local aesthetic-related information. The framework further introduces a Modal-adaptive Integration mechanism to combine visual and textual representations, allowing the same architecture to support different IAA modalities and task settings.

General-purpose MLLMs.

mPLUG-Owl2 (Ye et al., 2024). mPLUG-Owl2 adopts a modular multimodal architecture consisting of a vision encoder, a visual abstractor, a text embedding layer, and a language decoder. The language decoder serves as a universal interface for different modalities, while a modality-adaptive module projects modality-specific features into a shared semantic space and preserves modality-specific information. Its training paradigm jointly exploits unimodal and multimodal data to improve modality collaboration.

ShareGPT-4V (Chen et al., 2024). ShareGPT-4V is a general-purpose vision-language model trained with high-quality GPT-4V-generated image descriptions. Its multimodal architecture follows the LLaVA-style design, combining a pretrained vision encoder, an MLP projector, and a large language model. The ShareGPT-4V data provide detailed descriptions of visual objects, attributes, spatial relationships, world knowledge, and other image semantics, improving the visual-language alignment of the resulting model.

Qwen2.5-VL-7B (Bai et al., 2025b). See the previous subsection for details.

InternVL3-8B (Zhu et al., 2025). See the previous subsection for details.

Specialized IAA MLLMs.

Q-Instruct (Wu et al., 2024). Q-Instruct is a multimodal instruction-tuning approach designed to improve fine-grained low-level visual perception. It introduces the Q-Pathway dataset, which collects natural-language human feedback on perceptual properties such as clarity, color, brightness, and other low-level visual characteristics. These data are converted into visual instruction-tuning examples, enabling MLLMs to reason over detailed perceptual concepts beyond conventional numerical quality scores.

PEAS (Yun & Choo, 2024). PEAS addresses personalized image aesthetic assessment through task-vector customization. Rather than relying on a single personalized training database, it treats models fine-tuned on different image quality and aesthetic assessment datasets as task-specific vectors that encode distinct aesthetic characteristics. A personalized model is then obtained by learning a com-

bination of these task vectors from a small number of user-provided examples, enabling scalable adaptation to individual aesthetic preferences.

Q-Align (Wu et al., 2023). See the previous subsection for details.

UNIAA-LLaVA (Zhou et al., 2024). UNIAA-LLaVA is an MLLM-based unified IAA model built on the LLaVA architecture. Its key idea is to transform existing aesthetic assessment datasets with heterogeneous annotation formats into a unified visual instruction-tuning representation. This formulation allows the model to learn common aesthetic perception, description, and assessment capabilities across multiple datasets within a single multimodal framework.

Next Token Is Enough (Li et al., 2025). Next Token Is Enough investigates numerical aesthetic scoring using the native autoregressive generation capability of MLLMs. Instead of introducing a separate regression head, it represents numerical scores through next-token prediction and shows that predicting additional significant digits can substantially improve scoring accuracy. The approach further combines chain-of-thought reasoning with fine-grained aesthetic attributes to obtain more interpretable and accurate aesthetic predictions.

ArtiMuse (Cao et al., 2026). See the previous subsection for details.

Evaluation Datasets.

AVA (Murray et al., 2012). AVA is one of the largest publicly available image aesthetic assessment datasets, containing more than 250K images collected from the DPChallenge photography community. Each image is associated with a distribution of aesthetic ratings from multiple users, together with semantic labels covering more than 60 categories and additional photographic-style annotations. The ratings are provided on a 1–10 scale and reflect the perceived aesthetic quality of the images in the context of photography competitions. Owing to its large scale and diverse content, AVA is commonly used as a general-purpose benchmark for image aesthetic assessment.

PARA (Yang et al., 2022). PARA is designed to study the personalized nature of aesthetic perception. It contains 31,220 images annotated by 438 subjects, with each image receiving annotations across multiple dimensions. In addition to nine image-oriented objective attributes, PARA provides four human-oriented subjective attributes and anonymized subject information, enabling aesthetic preferences to be analyzed together with individual characteristics. The dataset therefore captures both image-level aesthetic properties and variations in aesthetic judgments across different subjects.

TAD66K (He et al., 2022). TAD66K focuses on the influence of semantic themes on aesthetic judgment. It contains approximately 66K images distributed across 47 popular themes, with the images manually selected according to their corresponding themes. In addition to common aesthetic criteria, TAD66K defines theme-specific evaluation criteria for the 47 themes, and each image is densely annotated by more than 1,200 annotators. This design explicitly accounts for the fact that aesthetic preferences may depend on the semantic context of an image.

FLICKR-AES (Ren et al., 2017). FLICKR-AES is a large-scale personalized aesthetics dataset constructed from photographs collected from Flickr. It contains 40,000 images annotated by 210 Amazon Mechanical Turk workers, with each image rated by five different workers on a 1–5 aesthetic scale. The average of the five ratings is used as the ground-truth aesthetic score. The dataset also preserves worker identities and adopts a user-disjoint train/test split, allowing evaluation of aesthetic prediction under unseen-user settings and providing a basis for studying variations in individual aesthetic preferences.

ArtiMuse-10K (Cao et al., 2026). ArtiMuse-10K is an expert-curated image aesthetic dataset containing 10,000 images spanning five major categories and 15 subcategories. Unlike datasets relying primarily on crowd-sourced holistic ratings, ArtiMuse-10K is annotated by professional experts using eight fine-grained aesthetic attributes together with a holistic aesthetic score. Its categories cover a broad range of visual content, including photography, painting and calligraphy, AIGC, 3D design, and graphic design. The resulting annotations provide both quantitative aesthetic judgments and fine-grained attribute-level assessments.

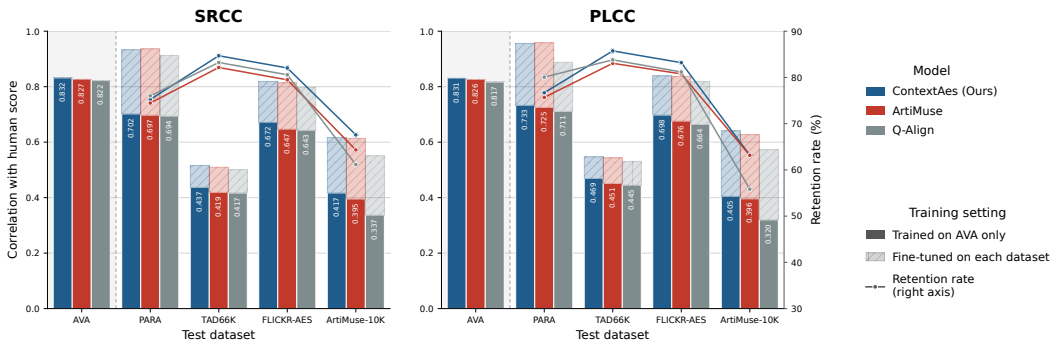


Figure 8: ContextAes-40K.

E.2.2 Results and Analysis

The complete cross-dataset results are shown in Figure 7. Performance varies substantially across datasets: PARA generally yields higher correlations, while TAD66K is consistently more challenging. FLICKR-AES shows intermediate difficulty, whereas ArtiMuse-10K is particularly challenging for general-purpose MLLMs, indicating that aesthetic prediction is sensitive to differences in image and annotation distributions.

Clear differences are also observed across model families. Traditional IAA methods can perform strongly on specific datasets but show less consistent behavior across benchmarks. General-purpose MLLMs generally achieve weaker aesthetic correlations, with Qwen2.5-VL-7B performing best within this group. Specialized IAA MLLMs achieve substantially stronger results, particularly Q-Align and ArtiMuse.

Among the three strongest specialized models, ContextAes performs competitively across all five datasets. ArtiMuse performs slightly better on PARA, whereas ContextAes achieves higher correlations on AVA, TAD66K, FLICKR-AES, and ArtiMuse-10K. The gap is more evident on the latter datasets, where ContextAes maintains stronger performance under different aesthetic annotation settings.

Figure 8 further analyzes cross-dataset robustness by training Q-Align, ArtiMuse, and ContextAes only on AVA and measuring their performance on the four unseen datasets. We normalize each target-domain result by the corresponding AVA performance to quantify the proportion of source-domain capability retained after distribution transfer. All three models experience performance degradation after transfer, with relatively high retention on PARA and FLICKR-AES and larger drops on TAD66K and ArtiMuse-10K. ContextAes shows higher retention on most unseen datasets. In particular, it retains 80.8% SRCC and 84.0% PLCC on FLICKR-AES, and 50.1% SRCC and 48.7% PLCC on ArtiMuse-10K, compared with 78.2%/81.3% and 78.2%/81.8% for Q-Align and ArtiMuse, respectively.

Overall, the retention analysis shows that source-domain performance does not fully reflect cross-dataset robustness. Although the three models achieve similar retention on PARA, ContextAes preserves a larger proportion of its AVA performance on most of the remaining datasets, suggesting more consistent transfer across different aesthetic evaluation distributions.

E.3 Ablation Study

The ablation results in Table 3 show that the effects of the three training stages are not simply additive. SFT alone achieves 74.27% overall Accuracy, whereas SkillOpt alone reaches 72.19%. When the two stages are combined, the overall Accuracy changes only slightly to 74.29%, suggesting that the benefit of SkillOpt is limited when used without subsequent preference optimization.

The effect of GRPO depends strongly on the preceding training stages. Adding GRPO to either SkillOpt or SFT alone reduces performance, with overall Accuracy dropping to 70.52% and 72.11%, respectively. In contrast, when SkillOpt and SFT are both applied before GRPO, the overall Accuracy and F1 increase to 75.30% and 75.27%, corresponding to gains of 1.01 and 1.12 percentage

points over SkillOpt+SFT. This difference suggests that GRPO benefits from the task-specific skill and supervised initialization established by the earlier stages.

The improvements from the complete training strategy are also more pronounced on the Editing subset. Compared with SkillOpt+SFT, the full strategy improves Editing Accuracy and F1 by 1.43 and 1.47 percentage points, respectively, while the corresponding gains on Natural are 0.66 and 0.78 percentage points. This indicates that the later preference optimization stage contributes more strongly when the collections contain controlled editing variations.

Across all ablation settings, Editing consistently achieves higher scores than Natural, with the gap increasing from roughly 5 percentage points in the simpler configurations to over 6 percentage points under the full training strategy. For the complete model, the Editing subset reaches 78.71% Accuracy and 78.60% F1, compared with 72.44% and 72.44% on Natural. The consistent difference across ablations suggests that the two collection types exhibit distinct levels of difficulty for aesthetic preference prediction.

E.4 Real-World Evaluation

Following the ranking-to-pairwise conversion described in the main paper, the 494 user-curated collections yield 26,688 pairwise test instances after exhaustive pairing and label-reversed augmentation. This construction preserves the relative ordering provided by each album provider while balancing the presentation order of the compared images, resulting in a unified test set for evaluating pairwise aesthetic preference prediction.

The results in Table 4 show that most general-purpose MLLMs achieve relatively similar performance, with Accuracy ranging from 55.87% to 59.13%. Model scale alone does not lead to consistently higher performance: InternVL3.5-30B performs below several smaller models, while Qwen3-VL-8B and Qwen3.5-9B obtain nearly identical Accuracy. Closed-source models also remain within a similar range, with Gemini3.1-Pro reaching 59.13% Accuracy, indicating that general multimodal capability alone does not fully capture the aesthetic preferences reflected in real-world album rankings.

Task-specific IAA models show greater variation. ArtiMuse reaches 57.34% on both Accuracy and F1, while FGAes and Q-Align obtain lower performance. In contrast, ContextAes achieves 63.64% Accuracy and 63.54% F1, exceeding the strongest Accuracy among the compared models by 4.51 percentage points and the strongest F1 by 4.81 percentage points. The relatively large gap suggests that the preference modeling learned from contextually related collections transfers beyond the controlled evaluation setting of ContextAes-40K.

The close correspondence between Accuracy and F1 for ContextAes also indicates stable pairwise predictions on the real-world test set. More broadly, the results show that the external album collections remain challenging for existing models, while ContextAes maintains a clear performance margin despite being evaluated on independently collected user-curated data.

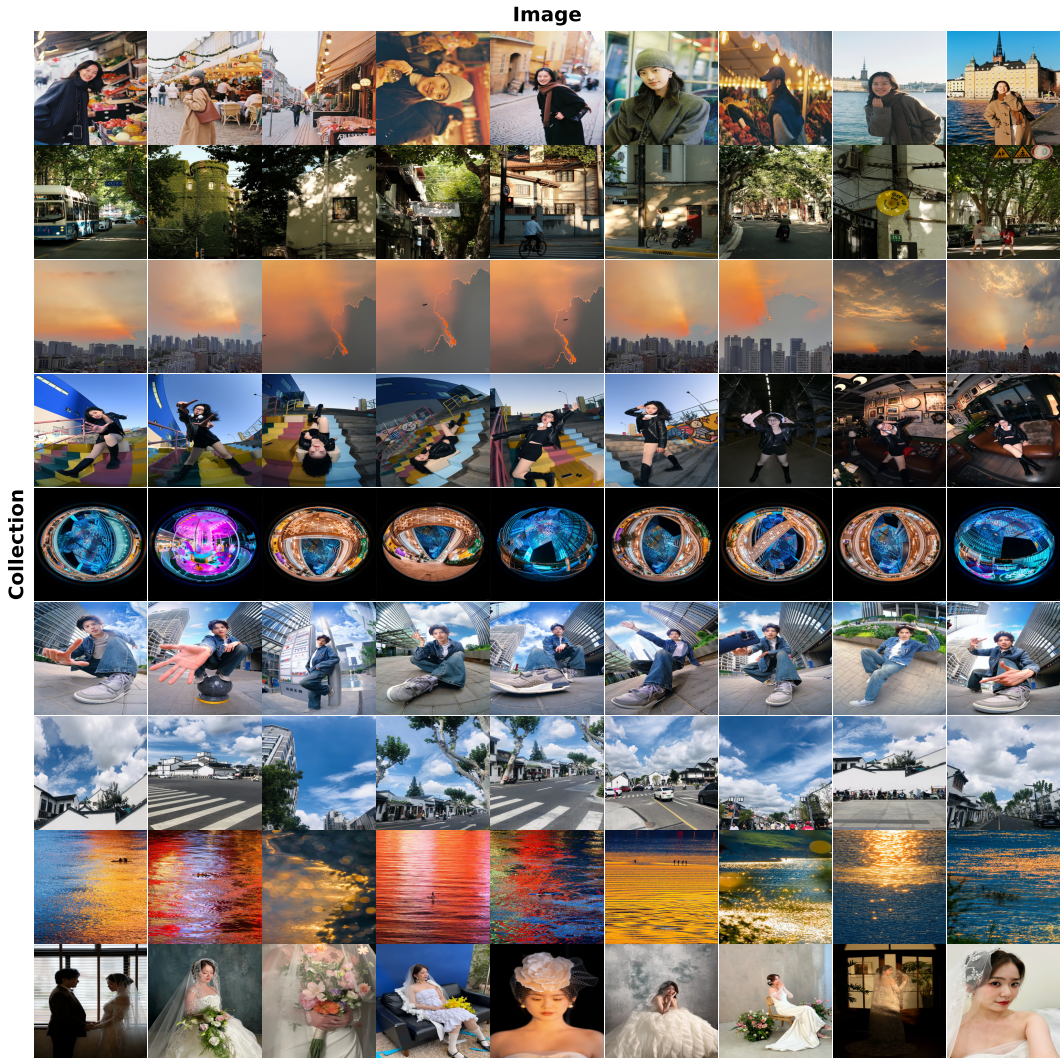


Figure 9: Examples of contextually related image collections from the Natural part of ContextAes-40K, collected from RedNote.



Figure 10: Examples of contextually related image collections from the Natural part of ContextAes-40K, collected from RedNote.

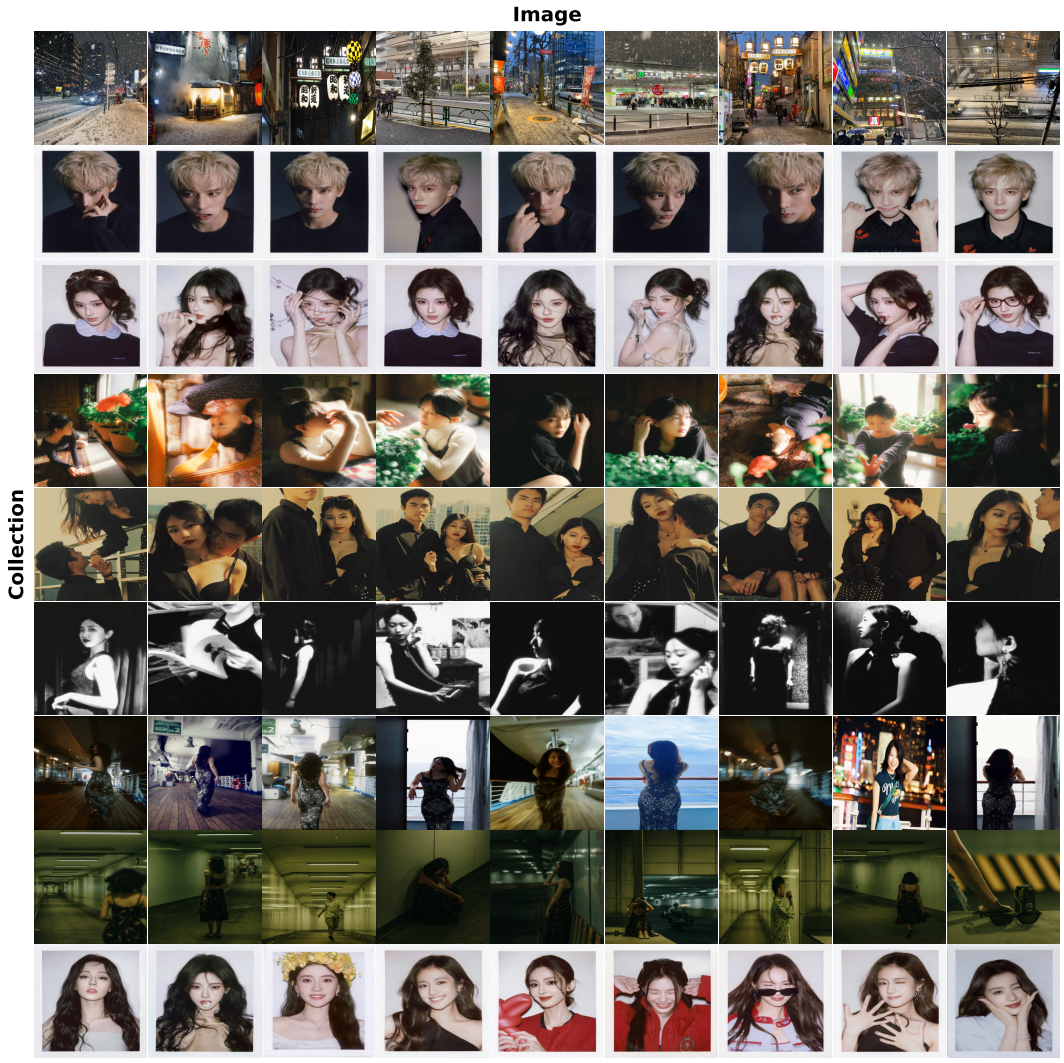


Figure 11: Examples of contextually related image collections from the Natural part of ContextAes-40K, collected from RedNote.

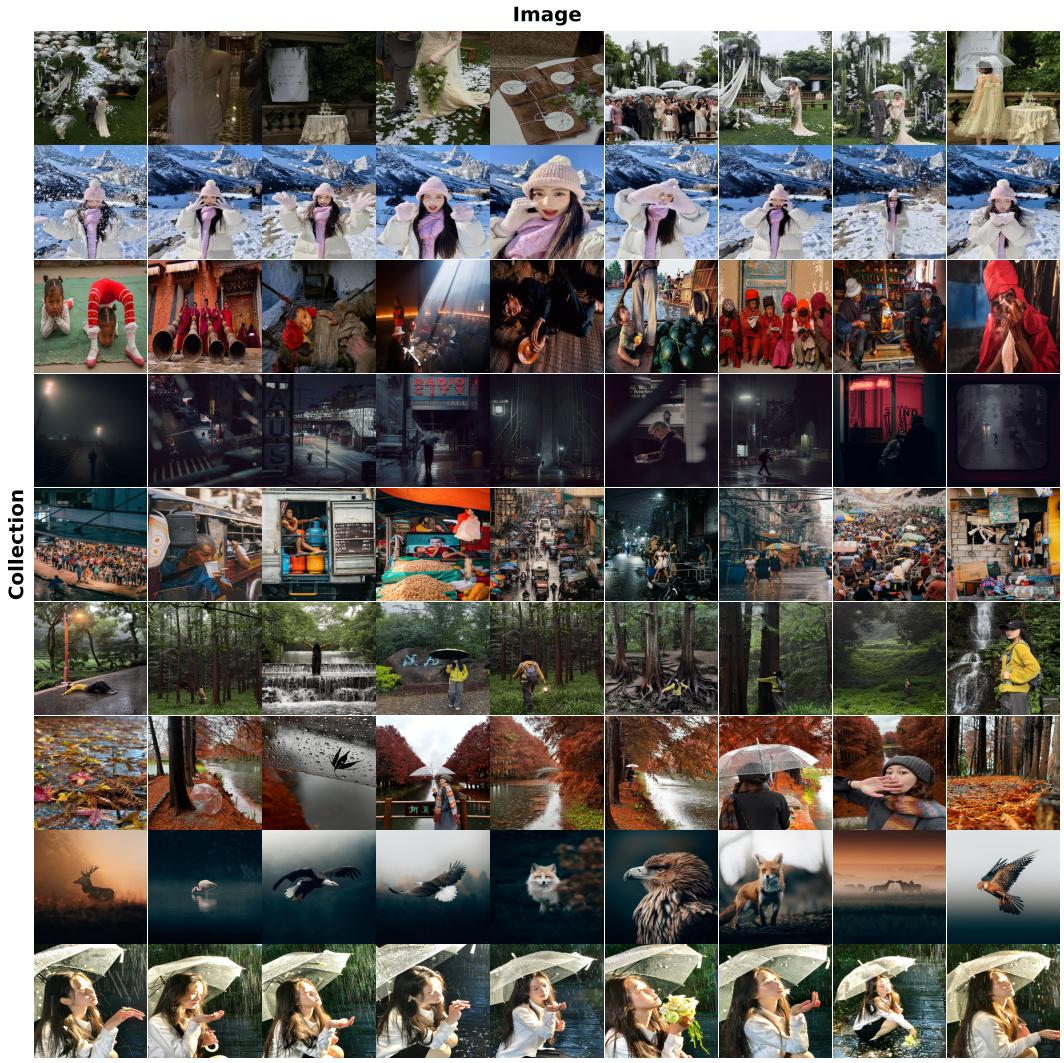


Figure 12: Examples of contextually related image collections in the Natural subset of ContextAes-40K collected from RedNote.

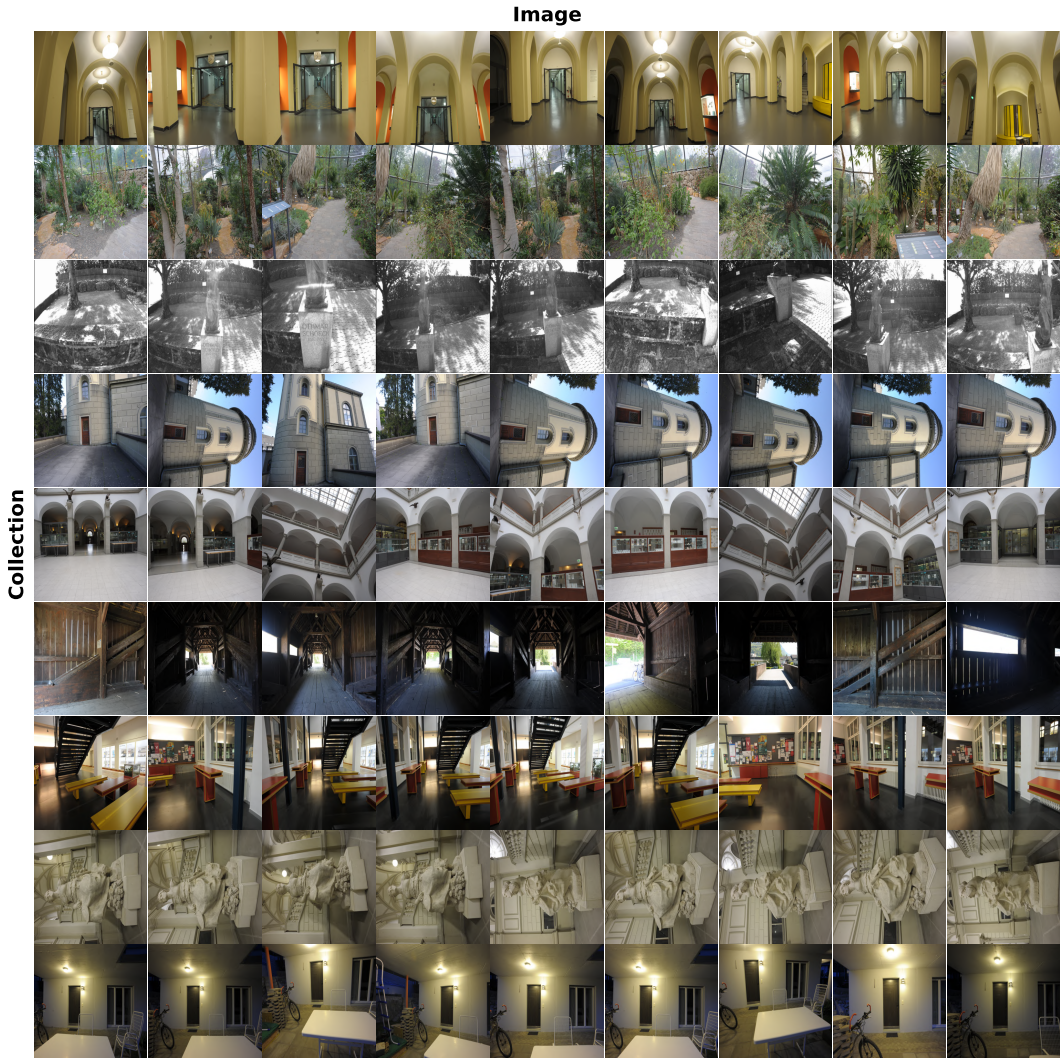


Figure 13: Examples of contextually related image collections from the Natural part of ContextAes-40K, collected from ETH3D.

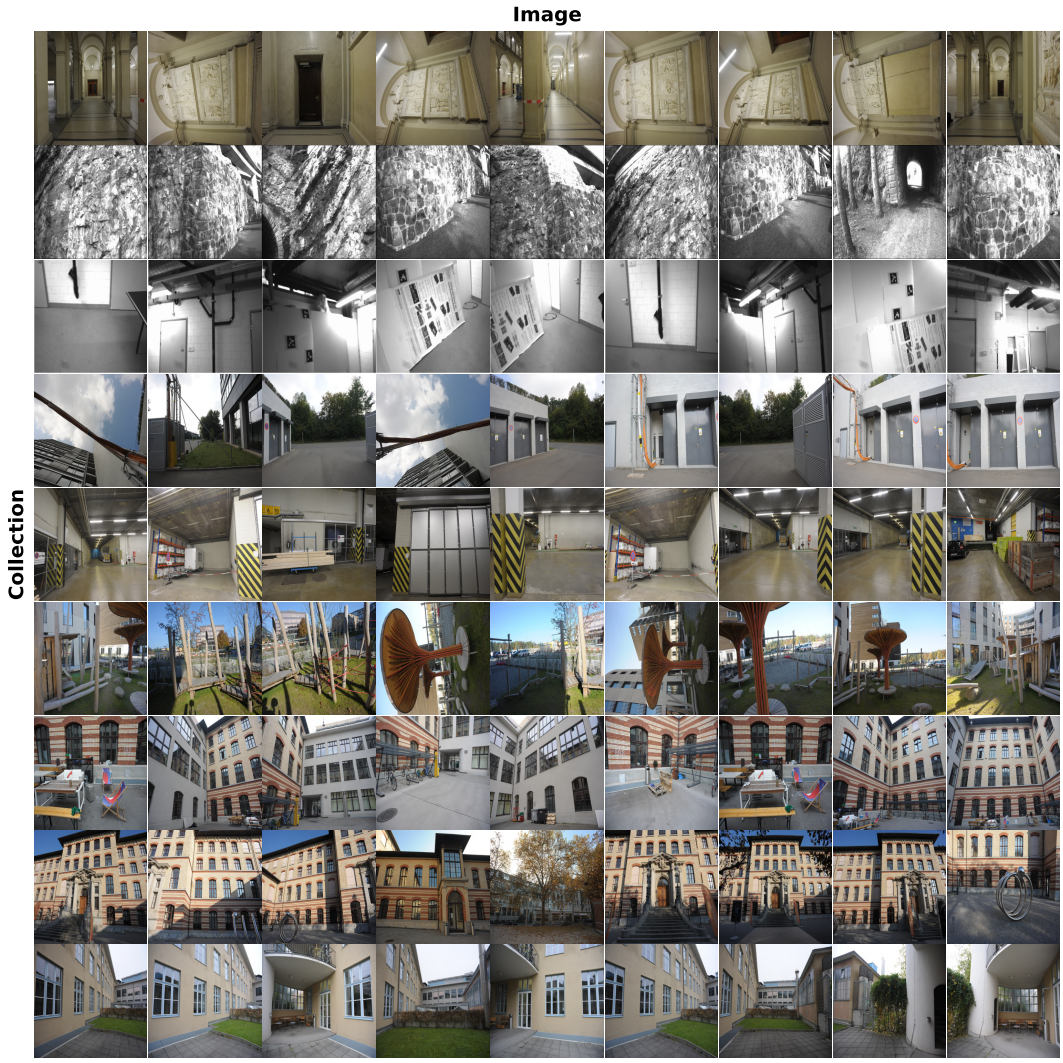


Figure 14: Examples of contextually related image collections from the Natural part of ContextAes-40K, collected from ETH3D.

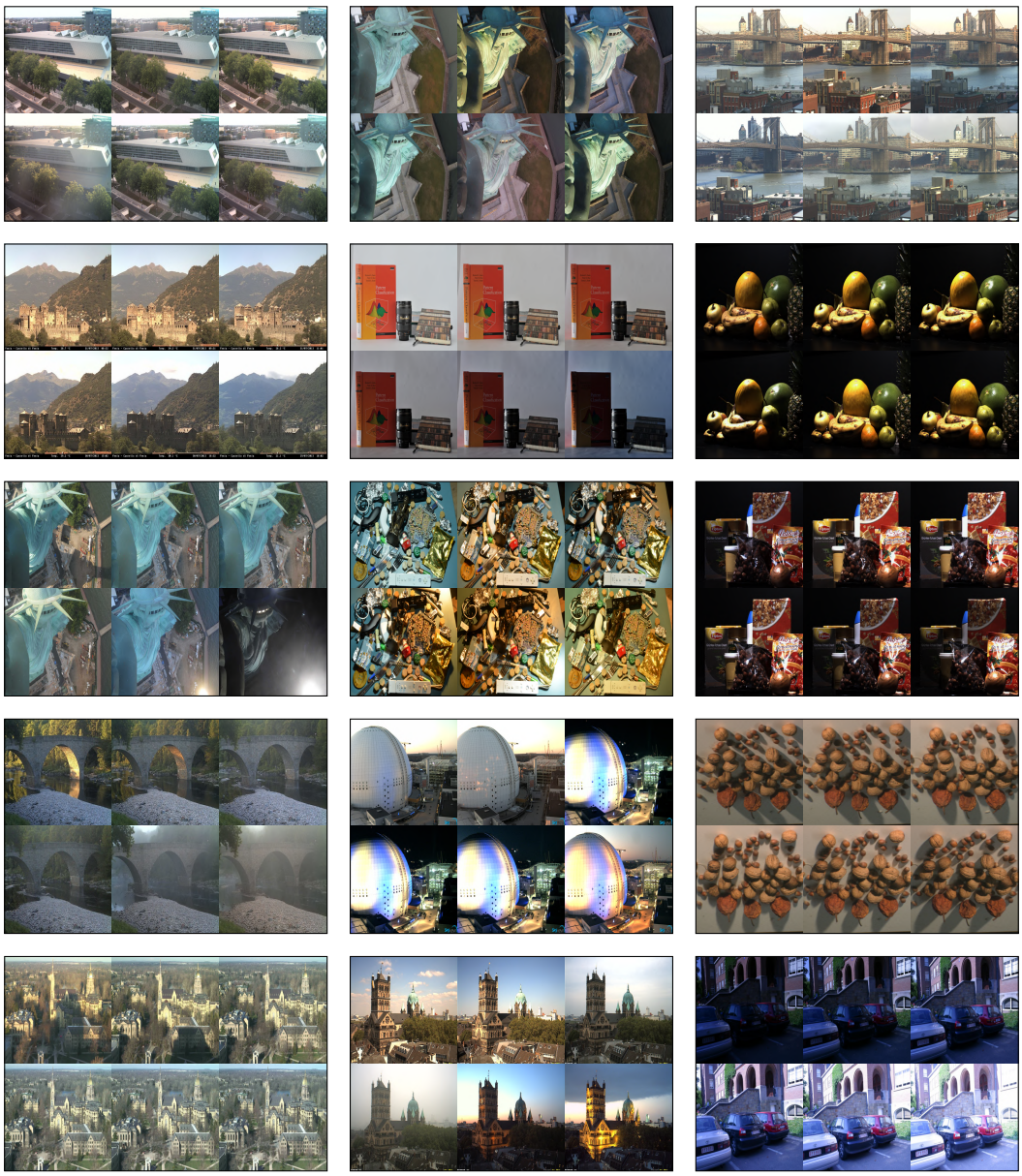


Figure 15: Examples of contextually related image collections from the Natural part of ContextAes-40K, collected from Hpatches.

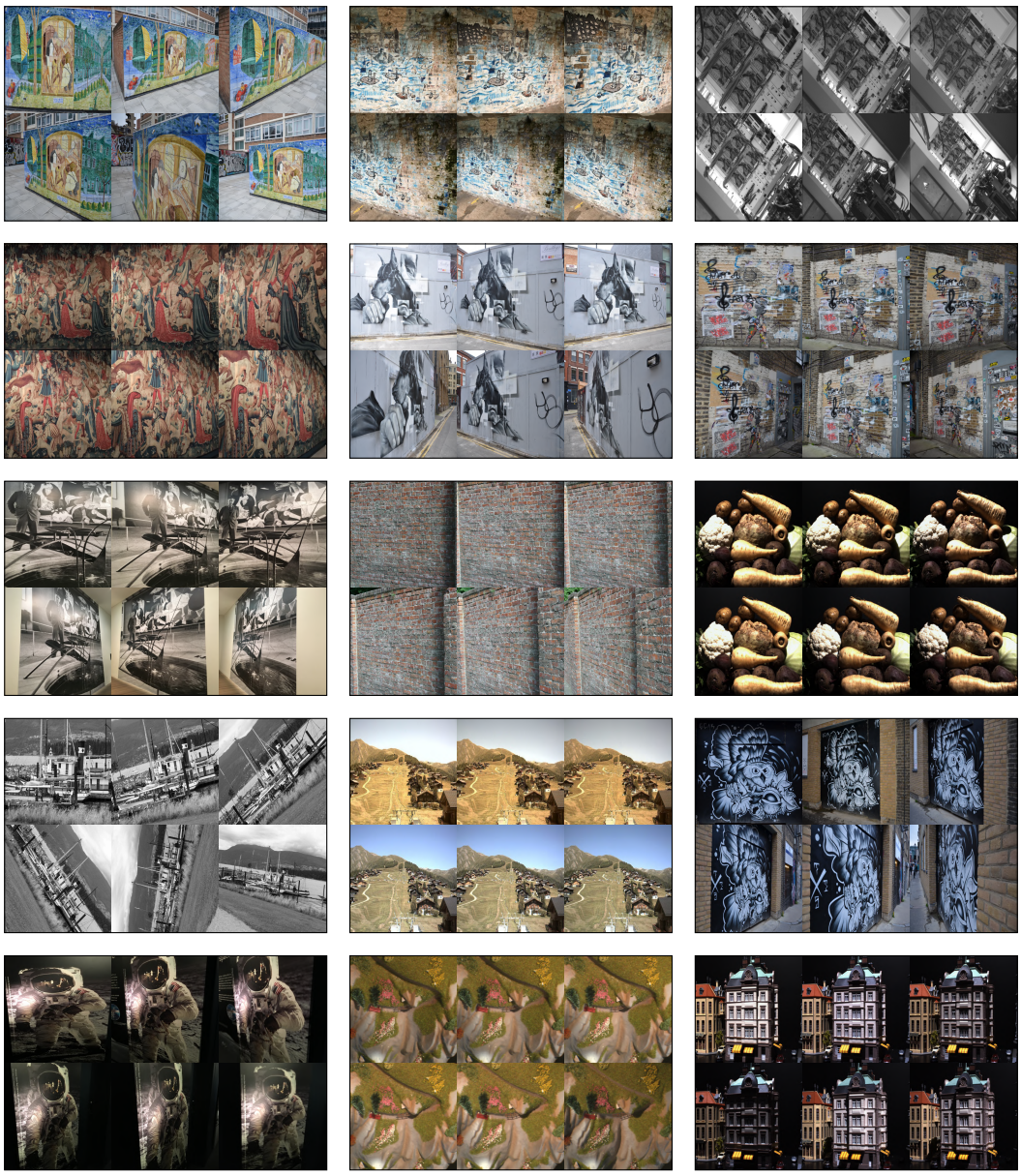


Figure 16: Examples of contextually related image collections from the Natural part of ContextAes-40K, collected from Hpatches.

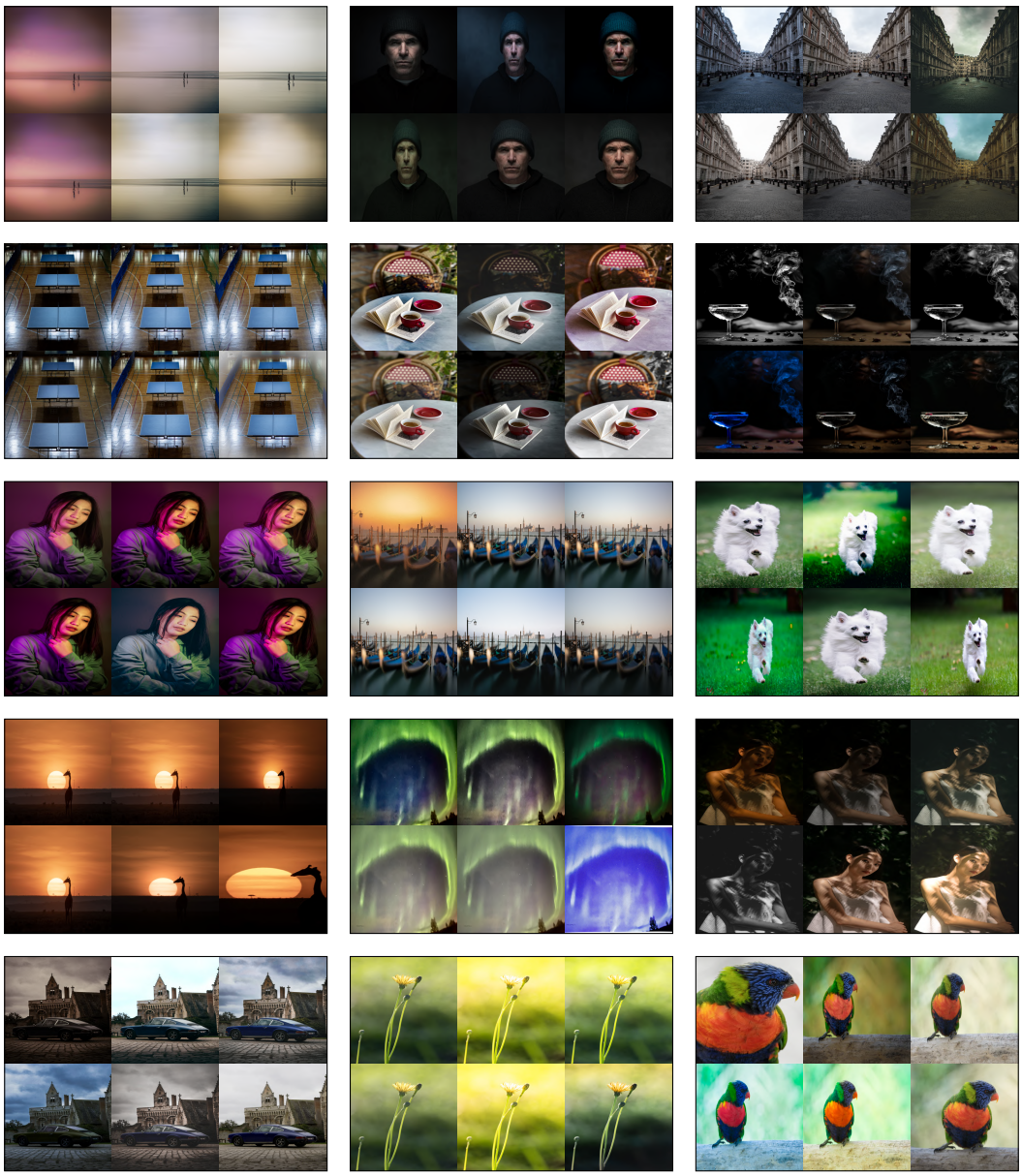


Figure 17: Examples of contextually related image collections from the Editing part of ContextAes-40K, collected from Lightroom.

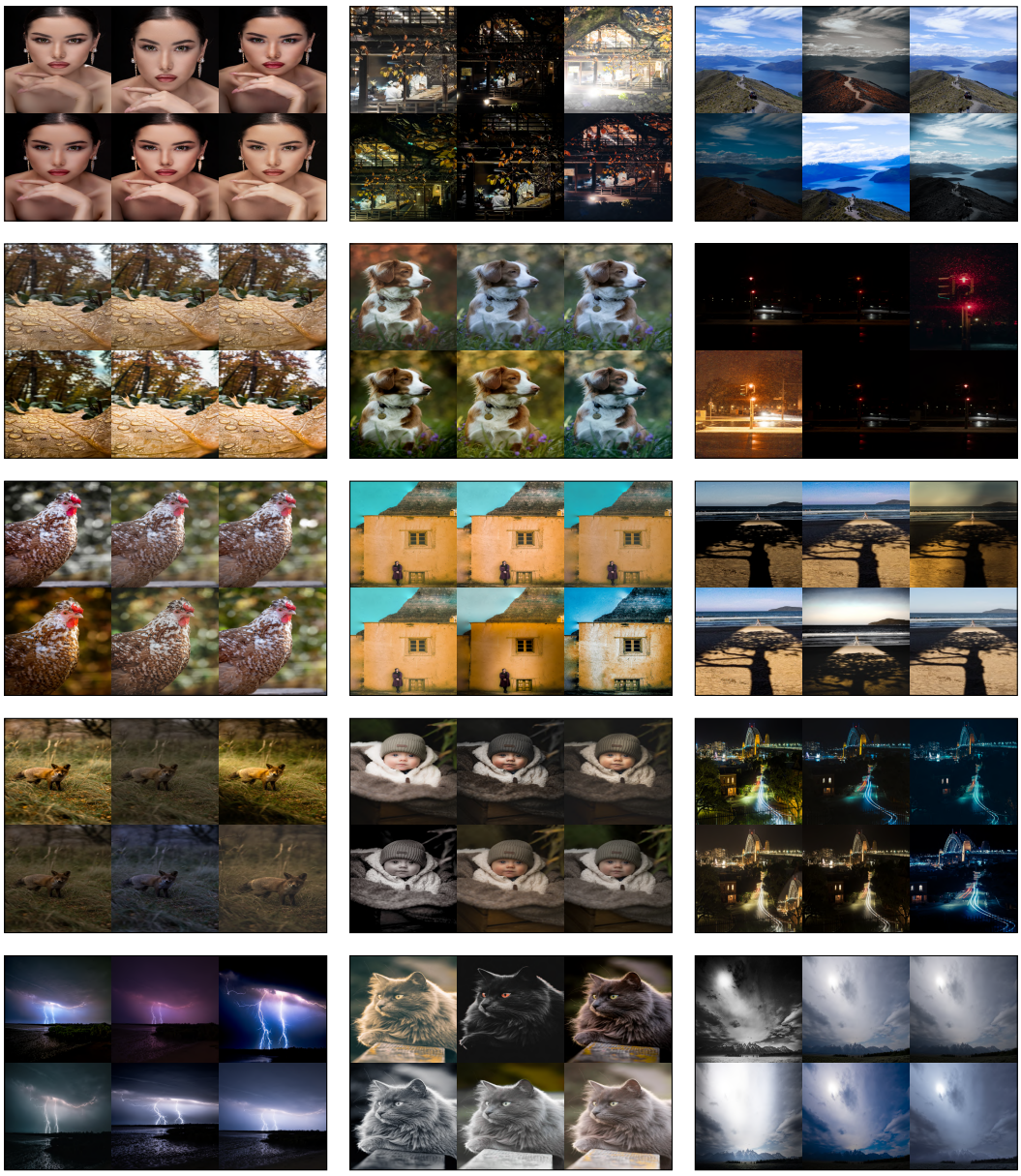


Figure 18: Examples of contextually related image collections from the Editing part of ContextAes-40K, collected from Lightroom.

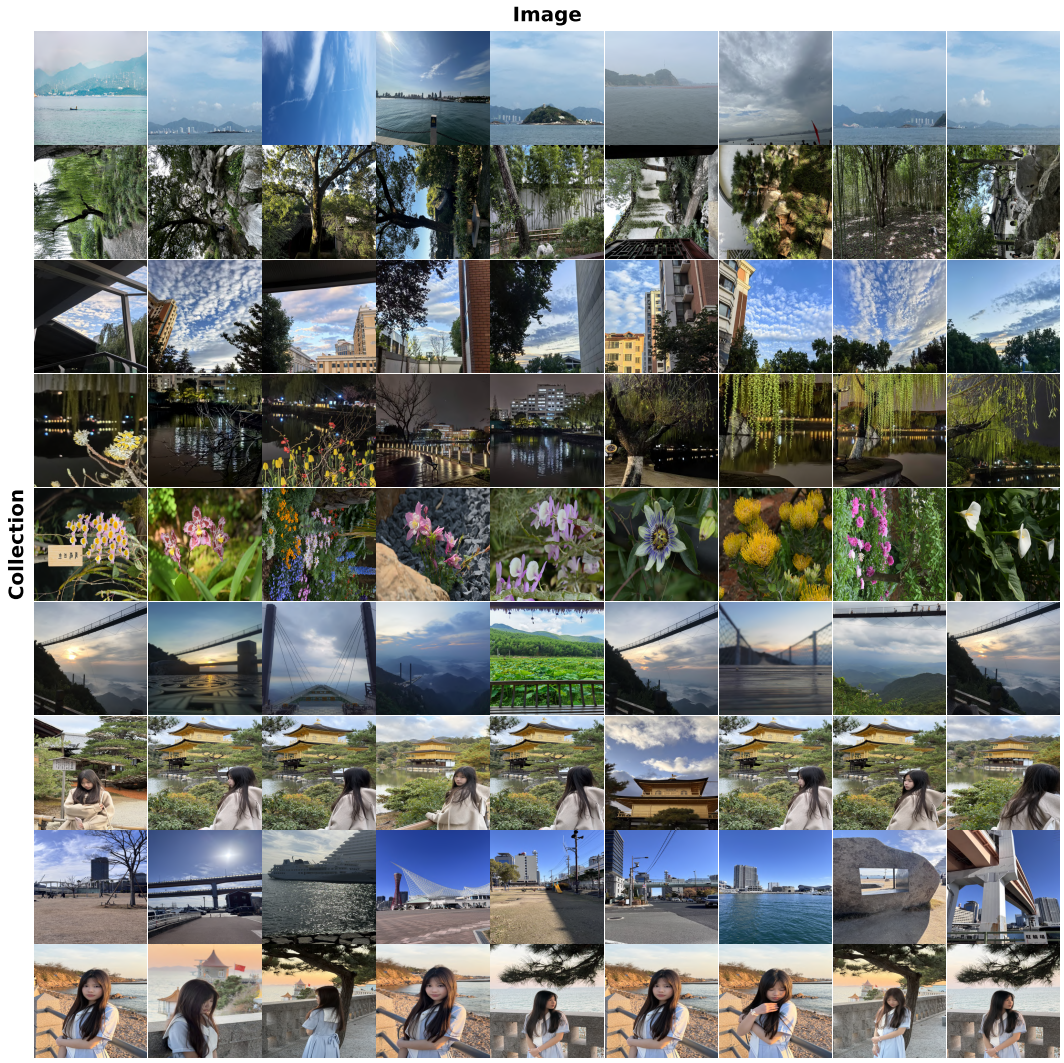


Figure 19: Examples of real-world user-curated photo collections from ContextAes-Album, with the top five images ranked according to users' aesthetic preferences.

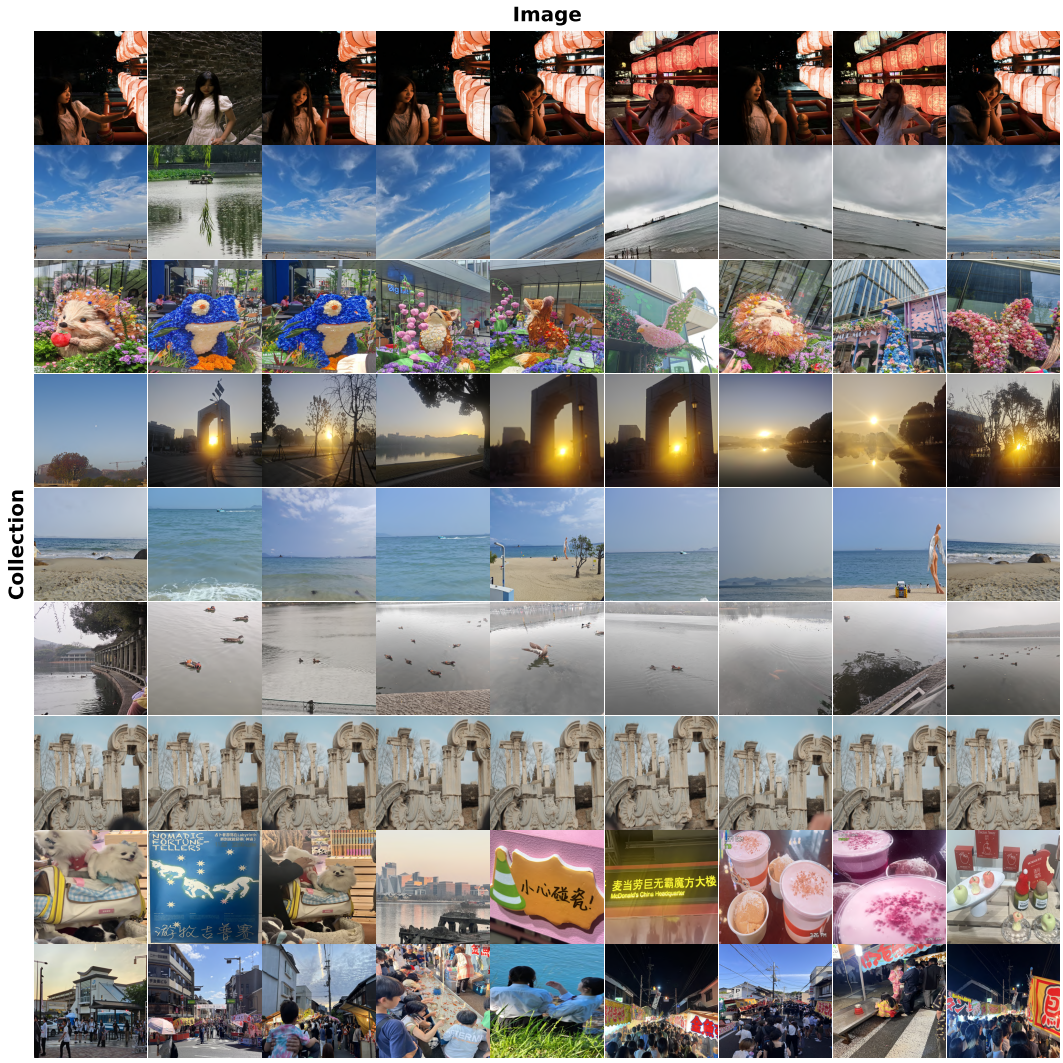


Figure 20: Examples of real-world user-curated photo collections from ContextAes-Album, with the top five images ranked according to users' aesthetic preferences.