AesX: Enhance Your Images with Stunning Aesthetic Beauty

Yuyan Chen[1], **Zhendong Hou**[2], **Lei Xia**[3], **Jiahao Li**[4], **Zhuolin Ji**[5], **Zhixu Li**[6] ✉
[1]Cornell University, [2]University of California, Los Angeles, [3]Tongji University
[4]Duke University, [5]Illinois Institute of Technology
[6]School of Information, Renmin University of China
yolandachen0313@gmail.com, dong0016@ucla.edu,
2210908@tongji.edu.cn, jiahao.li3@duke.edu,
zji2@hawk.iit.edu, zhixuli@ruc.edu.cn

Abstract

In the fields of advertising design, artistic creation, and cultural dissemination, there is an increasingly urgent demand for high-quality images that cater to fine-grained aesthetic preferences. Although existing large-scale models can generally meet basic requirements for clarity and alignment with textual elements, they still face significant bottlenecks in achieving precise control and aesthetic optimization. To address this limitation, we propose a set of comprehensive preference indicators across two major dimensions, text-image consistency and aesthetic quality, encompassing multiple criteria ranging from exposure and clarity to visual guidance and innovativeness. Building on these indicators, we have developed a generative framework named AesX to steer the model consistently toward a generation path that more closely aligns with human aesthetic sensibilities. Our experimental findings demonstrate that this approach yields significant improvements in both target recognition accuracy and overall visual aesthetic presentation.

1 Introduction

Image aesthetic quality that aligns with human preference is critically important and can play a significant role across a variety of fields and industries, such as advertising design, fashion e-commerce, and medical imaging. Current image generation models already demonstrate strong capabilities in producing high-quality images. For example, some recent works align with human preferences and achieve high aesthetic standards in aspects like text-image consistency and image clarity (Huang et al., 2024a,b), color richness (He et al., 2022). They design aesthetic scores to align with human preference (Li et al., 2024; Redies et al., 2024; Xiao et al., 2024). However, these existing approaches generally do not analyze human preference at a finer granularity, which leads to two primary problems: first, the generated images are often not of sufficiently high quality, and second, there is limited controllability over the quality attributes of these images. For example, while images that exhibit better text-image consistency by accurately depicting the described elements as shown in Fig. 1(b_1) compared with Fig. 1(a), our desired outputs require more refined aesthetic attributes such as balanced composition and emotional resonance as illustrated Fig. 1(c). Compared with Fig. 1(b_2), Fig. 1(c) demonstrates better aesthetic effects in terms of lighting and emotion.

We posit that one key solution to these two problems lies in designing more refined human preferences. Therefore, in this work, we first propose a set of metrics that encompass both text-image consistency and an aesthetic quality score. text-image consistency is defined through image subject accuracy and image subject completeness; the aesthetic quality score comprises exposure, clarity, color, noise control, focusing, rules usage, visual guidance, simplicity, sense of balance, unique perspective, creative expression, storytelling, emotional conveyance, authenticity, resonance, aesthetic value, cultural connotation, era significance, and innovativeness.

A second key solution is to integrate a knowledge enhancement module into the image generation process, such that these refined human preferences can be embedded within a Monte Carlo–based approach. Concretely, we propose a high-quality score image generation framework named AesX where Aes stands for aesthetic quality and X stands for extensibility [1]. AesX injects the aforementioned metrics as explicit knowledge instructions into the image generation model, thereby guiding it to produce images that more precisely satisfy these fine-grained human preferences.

A third solution is to introduce a preference enhancement module that provides real-time discrimination and control during the image generation

[1]https://github.com/Yukyin/AesX

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 2001–2011

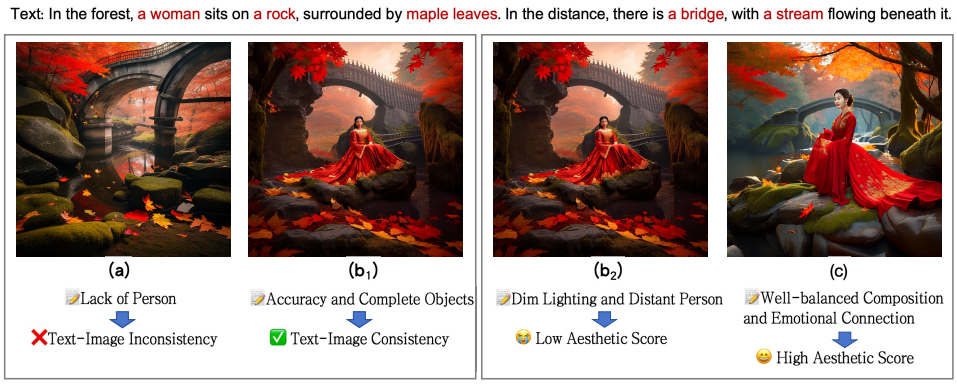

Figure 1: Comparing (a)'s text-image inconsistency with (b_1)'s basic accuracy, as well as (b_2)'s low aesthetic quality with (c)'s refined composition and emotional resonance.

process. Specifically, at each step of image generation, we integrate a reward model that performs a trade-off between depth and breadth. By comparing the newly generated image with the previously generated one and incorporating a referee mechanism, the model is incentivized to follow an optimal path toward producing higher-quality images.

2 Related Work

Text-to-Image Generation. Liang et al. (2024) introduced more extensive human-feedback mechanisms to evaluate and optimize the interpretability of text-to-image generation. Shi et al. (2024) utilized existing pretrained models to achieve personalized text-to-image generation without fine-tuning, thereby substantially enhancing efficiency. Wang et al. (2024c) designed multimodal prompts and post-processing mechanisms in text-to-image generation. Qin et al. (2024) constructed a language-model–driven text-to-image generation system for enhancing semantic expansion and cross-model interoperability. Wang et al. (2024a) introduced a cost-free method for preserving visual styles in text-to-image outputs which requires no extra training. Liao et al. (2024) proposed a specialized cross-modal framework for abstract concepts. Zhou et al. (2024b) introduced a plug-and-play customization assistant for rapid adaptation in large models without parameter tuning. Although these works have made significant progress in text-to-image generation, and some align with human preferences, there is still considerable room for improvement in generating images that meet human aesthetic standards.

Image Aesthetics Assessment. Huang et al. (2024a) proposed a multimodal foundation model for image aesthetic perception. Huang et al. (2024b) constructed an expert benchmark for multimodal large-scale models and explored the limitations and potential of such models in capturing image aesthetic perception from multiple perspectives. Chen et al. (2024) incorporated emotional information and a multibranch network to predict the aesthetic distribution. Yan et al. (2024) presented a hybrid CNN-Transformer–based meta-learning approach for personalized image aesthetics assessment. Soydaner and Wagemans (2024) employed a multi-task convolutional neural network for image aesthetic assessment. Jia et al. (2025) adopted a self-supervised strategy to incorporate Transformers into image aesthetic assessment tasks. Jia et al. (2022) proposed a no-reference image quality assessment method via non-local dependency modeling, further advancing blind quality estimation without relying on reference images. Xiao et al. (2024) introduced a multi-granularity fusion network to boost multimodal understanding by focusing on both image aesthetics and sentiment analysis. Mo et al. (2024) proposed a dynamic prompt-optimization strategy to improve both the aesthetic performance of generated images and alignment with textual semantics. However, none of these approaches have identified fine-grained image aesthetics features and integrated them into heuristic reasoning paths to guide the model in enhancing the aesthetic quality of generated images.

3 AesBench and Image Aesthetic Score

3.1 AesBench

We construct a novel aesthetic image benchmark named AesBench which aims to evaluate the capabilities of text-to-image models in generating high aesthetic images. We first design the image generation task and then introduce specific metrics for high-quality images to quantitatively evaluate large-scale image generation models. Approximately 70,000 images are collected from GQA,

COCO, and ImageNet, from which target objects are extracted using GPT-4o. Only those images containing over five but no more than fifteen objects are retained, after which a dual validation process involving both a large-scale model and human annotators is conducted. Objects are not treated as a fixed closed-set label ontology. Instead, they are open-ended descriptions that capture the basic semantic composition of each image. To generate the object list, we adopt the object-extraction prompt proposed by Chen et al. (2025). Subsequently, GPT-4o is employed for self-validation to verify whether the textual descriptions fully and accurately capture all objects in each image, scoring completeness and accuracy as binary indicators (0 or 1). We also adopt human evaluation for further validation and ultimately obtain 5,000 entries. The details of validation are introduced in Sec. A in Appendix.

3.2 Image Aesthetic Score

The image aesthetic score consists of two key aspects, *Text-Image Consistency* and *Image Attribute Quality*. Text-image consistency comprises two components: image object accuracy, indicating whether each depicted object aligns with the textual description, and image object completeness, determining whether all mentioned objects are fully represented. The image attribute quality encompasses i) *Technical Quality*, reflecting appropriate exposure, clarity, color fidelity, noise control, and focusing; ii) *Composition Form*, assessing the use of compositional rules, visual guidance, simplicity, and overall balance; iii) *Creativity Expression*, examining distinctive perspectives, novel techniques, and storytelling elements; iv) *Emotion resonance*, evaluating how effectively the image conveys emotion and evokes viewer responses; and v) *Aesthetics Culture*, emphasizing artistic value, cultural context, temporal significance, and innovation.

Each image also receives a mean opinion score (MOS) in the continuous range [0, 100] to represent overall quality, assigned by three expert annotators and subdivided into five ordinal levels (bad, poor, fair, good, excellent). Additionally, each image is annotated with an accuracy score (0 or 1) and five aesthetic scores, each using the same continuous [0, 100] scale and corresponding five-level labels. Higher scores in these metrics indicate better perceived quality. Specifically, the MOS, accuracy score, and five aesthetic scores are all human-annotated. The MOS and each aesthetic score use a continuous [0, 100] scale subdivided into five ordinal levels (bad, poor, fair, good, excellent); the accuracy score is binary (0 or 1). To ensure the reliability of these human judgments, inter-rater agreement is measured using Krippendorff's Alpha (IRA), and any ratings with an IRA below 0.7 are deemed controversial and replaced with descriptions from other images.

4 Methods

In this section, we design a framework named AesX to enhance the aesthetic quality of images generated by text-to-image generation models. AesX comprises two modules: a knowledge enhancement module and a preference optimization module, as illustrated in Figure 2.

4.1 Knowledge Enhancement

The knowledge enhancement module integrates the above quantitative indicators into instructions to guide text-based image generation. Formally, given object set $\mathcal{O}$, the pipeline proceeds through three stages. The first stage converts the object list into coherent text $T_c = G_{\text{coh}}(\mathcal{O})$. The second stage performs text-side aesthetic expansion to obtain description-enhanced text $T_d = G_{\text{aes}}(T_c)$. The third stage combines the predefined high-aesthetic feature set $\mathcal{K}$ with T_d and the initial generated image I to produce the aesthetically refined text $T_a = G_{\text{ref}}(T_d, I, \mathcal{K})$.

In the first stage, we employ GPT-4o to convert the object list into coherent text, enabling the image generation model to understand relationships among objects and produce a conventionally acceptable image. For example, the original text is "In the forest, a woman sits on a rock, surrounded by maple leaves. In the distance, there is a bridge, with a stream flowing beneath it." The prompt used is: "Using the object list: {object_list}, write a scene description that naturally incorporates all the objects. Use simple sentences but do not make the sentences too short."

In the second stage, GPT-4o performs text-side aesthetic expansion on this coherent text, adding descriptive details to yield description-enhanced text such as "In an autumn forest, a woman wearing a Victorian-style gown sits on a moss-covered rock, surrounded by red and golden maple leaves..." This guides the model toward a coarse-grained aesthetic outcome. The prompt used is: "Based on the

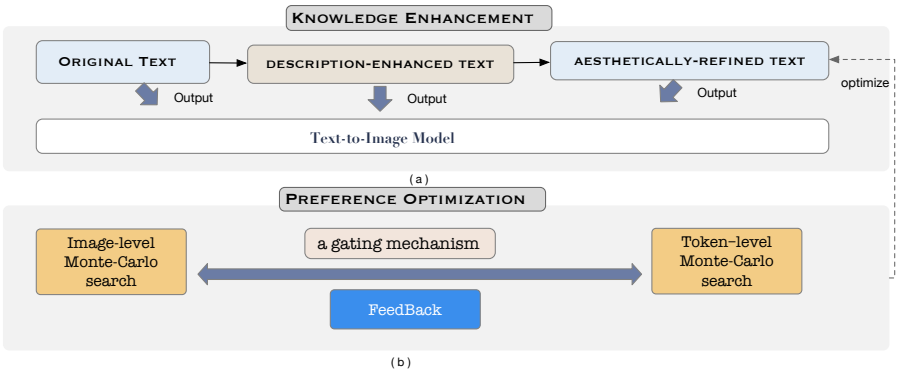

Figure 2: Framework of AesX.

above scene description, generate an aesthetically enhanced version for text-to-image generation. Preserve the original semantic content, object composition, and scene layout. Add rich and concrete visual details, such as season, colors, textures, materials, clothing style, architectural style, surrounding vegetation, light, and atmosphere. The output should feel vivid, elegant, and cinematic, while remaining faithful to the original scene." Both prompts were manually optimized through iterative prompt engineering.

In the third stage, after the model generates an initial image, we incorporate the predefined high-aesthetic features $\mathcal{K}$ into T_d to form the aesthetically refined text T_a, requiring the model to produce a higher-quality image satisfying these features and to provide per-feature explanations such as "Adjust the color to make the scene more vibrant..." Through these three stages, we generate images that align with fine-grained human aesthetic preferences.

4.2 Preference Optimization

The preference optimization module aims to select the higher aesthetic image from two candidates at each iteration and repeat this process to generate the highest aesthetic image. The evaluation criterion ensures that both images first satisfy text-image consistency before comparing their aesthetic scores. If one image fails the text-image consistency requirement, it is discarded immediately. At each generation step, GPT-4o serves as the reward model to provide feedback. GPT-4o evaluates both text-image consistency and aesthetics jointly, and the reward function $R(\cdot)$ maps GPT-4o feedback together with the current image and its corresponding instruction text to a continuous scalar reward score in $[0, 100]$, derived by aggregating the five aesthetic dimensions.

We first use the aesthetically-refined text T_a as the instruction and conduct a Monte Carlo search which balances breadth-first search (BFS) and depth-first search (DFS) at the coarse-grained image level to find better output paths. At node i, the current candidate image is denoted as I_i and the corresponding refined instruction text as T_i^a. AesX generates multiple complete candidate images and evaluates them based on GPT-4o feedback C_i on the current image I_i and instruction T_i^a. The image-level score is defined as:

$$S_{\text{image}}(I_i) = R(C_i, I_i, T_i^a) \tag{1}$$

Simultaneously, instead of evaluating isolated local patches, we introduce a CLS token representing the global state of the entire image, including overall composition, color harmony, style, and object relationships, such that token-level search still evaluates whole-image aesthetics. After generating the current candidate image I_i, we obtain GPT-4o feedback C_i^{cls} based on I_i and T_i^a, and define the token-level score as:

$$S_{\text{token}}(I_i) = R(C_i^{\text{cls}}, I_i, T_i^a) \tag{2}$$

Next, we design a gating mechanism to dynamically choose the iteration granularity at each step, introducing learnable weight parameters λ_1 and λ_2 to balance the depth and breadth of the image generation process:

$$S(I_i) = \lambda_1 S_{\text{image}}(I_i) + \lambda_2 S_{\text{token}}(I_i) \tag{3}$$

We then compare the fused score of the current node with that of its parent node and define a binary flag:

$$F_i = \begin{cases} 1, & \text{if } S(I_i) > S(I_{i-1}) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where F_i indicates whether the current candidate achieves a higher fused aesthetic score than its parent node. Each node n_i stores the image I_i, comments C_i, and flag F_i.

This strategy better satisfies fine-grained human aesthetic preferences, with every step being evaluated by GPT-4o feedback. The process terminates once no significant improvement is observed over the most recent α iterations:

$$\max_{i\in[1,\alpha]} \big(S(I_i) - S(I_{i-1})\big) \leq \epsilon \tag{5}$$

where ϵ is a predefined threshold determined through experiments. By iterating through these preference-driven optimizations, our model progressively refines the image to achieve the highest aesthetic quality while maintaining strict text-image consistency.

5 Experiments

In this section, we design experiments to demonstrate that the AesX effectively facilitates the generation of images that align more closely with fine-grained human preferences.

5.1 Experimental Setup

In this study, we conduct image generation experiments using eight A100 GPUs and a batch size of 8. During the Monte Carlo search process, α is set to 5, meaning that generation terminates if no superior image is produced in five consecutive steps. We set Top-k to 16,384, Top-p to 1.0, and temperature to 1.0 to accommodate diverse outputs. Emu3's output resolution is 512×512, and all other hyperparameters follow their respective model defaults unless otherwise specified.

The computational cost of AesX is primarily determined by the branching factor b and realized search depth d, with practical complexity of $O(b \cdot d)$. Since AesX only compares two candidate images per iteration and terminates early after α=5 non-improving steps, both b and d remain tightly bounded, preventing exponential growth. AesX is therefore better suited to high-quality offline generation scenarios rather than latency-sensitive real-time deployment.

5.2 Datasets, Baselines, and Metrics

We utilize SPAQ (Fang et al., 2020) and TAD66K (He et al., 2022) for image-aesthetics evaluation and adopt MSCOCO (Chen et al., 2015), GenEval (Ghosh et al., 2023), T2I-CompBench (Huang et al., 2023) and DPG-Bench (Hu et al., 2024) as general text-to-image benchmarks. Our baselines cover both autoregressive and diffusion models as shown in Table 3 with the details in Sec. B in Appendix. We use machine and aesthetic metrics for evaluation. Machine metrics are similar as the previous work which are also introduced in Sec. B in Appendix. For aesthetic evaluation, we utilize SRCC (Spearman rank correlation coefficient) to gauge the rank-order consistency between objective measures and human MOS (mean opinion score), which, alongside continuous attribute scores for brightness, color, contrast, noise, and sharpness, provides a more fine-grained measurement of image quality.

Table 1: Comparison of results in models before and after adding AesX in AesBench. TIC: Text-Image Consistency, TQ: Technical Quality, CF: Composition Form, CE: Creativity Expression, ER: Emotion Resonance, AC: Aesthetics Culture.

(AesBench) Method	Accuracy TIC	Aesthetics TQ	CF	CE	ER	AC	Total_A	Total
Diffusion-based								
SDXL	0.743	3.27	3.14	2.26	2.76	1.80	2.65	0.637
+AesX	0.76	3.41	3.27	2.29	2.88	1.83	2.74	0.654
PixArt-alpha	0.651	2.55	2.49	1.89	1.93	1.57	2.09	0.535
+AesX	0.674	2.75	2.63	1.92	1.96	1.62	2.18	0.555
DALLE3	0.865	4.15	4.14	3.15	3.29	2.24	3.39	0.772
+AesX	0.883	4.23	4.27	3.20	3.31	2.28	3.46	0.788
SD3	0.887	4.27	4.20	3.11	3.37	2.23	3.44	0.788
+AesX	0.890	4.28	4.33	3.17	3.39	2.26	3.49	0.794
Autoregressive-based								
Show-o	0.712	3.13	3.01	2.13	2.44	1.68	2.48	0.604
+AesX	0.742	3.46	3.35	2.24	2.48	1.74	2.65	0.636
Transfusion	0.783	3.56	3.55	2.34	2.77	1.98	2.84	0.676
+AesX	0.799	3.70	3.74	2.47	2.92	2.02	2.97	0.697
Chameleon	0.595	2.37	2.44	1.72	1.88	1.34	1.95	0.493
+AesX	0.613	2.74	2.71	1.83	1.94	1.41	2.13	0.520
LlamaGen	0.528	2.12	2.06	1.55	1.74	1.21	1.74	0.438
+AesX	0.572	2.62	2.54	1.70	1.80	1.30	1.99	0.485
Emu3	0.843	4.08	3.89	2.95	2.85	2.13	3.18	0.740
+AesX	0.872	4.22	4.14	3.06	3.01	2.18	3.32	0.768
Average	0.734	3.28	3.21	2.34	2.56	1.80	2.64	0.631
Average_A	0.756	3.49	3.44	2.43	2.63	1.85	2.77	0.655
↑	**0.022**	**0.21**	**0.23**	**0.09**	**0.07**	**0.05**	**0.13**	**0.024**
↑(%)	**2.997**	**6.40**	**7.17**	**3.85**	**2.73**	**2.78**	**4.59**	**1.958**

Table 2: The effect of AesX on Image Aesthetics Assessment datasets.

Method	SPAQ Accuracy	SPAQ Aesthetics	SPAQ Total	TAD66K Accuracy	TAD66K Aesthetics	TAD66K Total
DALLE3	0.895	3.83	2.363	0.883	3.64	2.262
+AesX	0.903	3.92	2.412	0.897	3.75	2.324
Emu3	0.871	3.36	2.116	0.855	3.49	2.173
+AesX	0.883	3.54	2.212	0.867	3.71	2.289
Average	0.883	3.60	2.242	0.869	3.57	2.220
Average_A	0.893	3.73	2.312	0.882	3.73	2.306
↑	**0.010**	**0.13**	**0.070**	**0.013**	**0.16**	**0.087**
↑(%)	**1.133**	**3.61**	**2.372**	**1.496**	**4.48**	**2.988**

Table 3: The effect of AesX on Text-to-Image datasets.

Method	MSCOCO			GenEval	T2I-CompBench			DPG-Bench
	CLIP-I	CLIP-T	FID	Overall	Color	Shape	Texture	Average
Diffusion-based								
SDv1.5	0.667	0.302	9.93	0.43	0.3730	0.3646	0.4219	63.18
DALLE2	-	0.314	10.93	0.52	0.5750	0.5464	0.6374	-
SDv2.1	-	-	-	0.50	0.5694	0.4495	0.4982	-
SDXL	0.674	0.310	-	0.55	0.6369	0.5408	0.5637	74.65
PixArt-alpha	-	-	7.32	0.48	0.6886	0.5582	0.7044	71.11
DALLE3	-	0.320	-	0.67	0.8110	0.6750	0.8070	83.50
DALLE3+AesX	-	**0.350**	-	**0.69**	**0.8201**	**0.6850**	**0.8132**	**84.42**
SD3	-	-	-	0.74	-	-	-	-
Autoregressive-based								
Emu	0.656	0.286	11.60	-	-	-	-	-
Show-o	-	-	9.24	0.53	-	-	-	-
Transfusion	-	-	6.78	0.63	-	-	-	-
Chameleon	-	-	26.74	0.39	-	-	-	-
LlamaGen	-	-	12.80	0.32	-	-	-	-
Emu3	0.689	0.313	12.80	0.66	0.7913	0.5846	0.7422	80.60
Emu3+DPO	0.68	0.312	19.30	0.64	0.7544	0.5706	0.7164	81.60
Emu3+AesX	**0.683**	**0.313**	**20.50**	**0.68**	**0.8125**	**0.6043**	**0.7601**	**82.31**

Table 4: The effect of each module in AesX.

	Accuracy	Aesthetics					Total_A	Total	↓	↓(%)
	TIC	TQ	CF	CE	ER	AC				
Emu3	0.843	4.08	3.89	2.95	2.85	2.13	3.18	0.740	**0.028**	**3.784**
w/o KnowE	0.85	4.10	3.98	2.98	2.91	2.14	3.22	0.747	**0.021**	**2.811**
w/o PreE	0.861	4.14	4.01	3.01	2.94	2.14	3.25	0.756	**0.012**	**1.587**
w/o image-level PreO	0.864	4.17	4.04	3.03	2.95	2.15	3.27	0.759	**0.009**	**1.186**
w/o token-level PreO	0.870	4.18	4.08	3.03	3.00	2.18	3.29	0.764	**0.004**	**0.524**
Emu3+AesX	0.872	4.22	4.14	3.06	3.01	2.18	3.32	0.768	-	-

5.3 Main Results

The results in Table 1 are obtained through machine evaluation. From Table 1, it is evident that all models exhibit improvements in both text-image consistency and aesthetic-related metrics after incorporating AesX. The average text-image consistency increases from 0.734 to 0.756 (a gain of about 2.997%), the technical quality metric rises from 3.28 to 3.49 (an increment of roughly 6.4%), and the aesthetic score moves up from 2.64 to 2.77 (an improvement of approximately 4.59%). For diffusion-based methods like DALLE3 and SD3, they already maintain high accuracy in target identification, yet AesX further enhances their performance in more fine-grained aesthetic dimensions, including composition and creative expression. For autoregressive models like Emu3, the preference-enhancement and multi-round feedback mechanisms further optimize core metrics: text-image consistency moves from 0.843 to 0.872, Average_A rises from 3.18 to 3.32, and the final average increases from 0.74 to 0.768.

From Table 2, it is evident that on the relatively simple SPAQ dataset (mainly featuring single-object scenarios), both the diffusion-based model exemplified by DALLE3 and the autoregressive model typified by Emu3 exhibit clear improvements in Accuracy and Aesthetics scores after the introduction of the AesX, suggesting that AesX effectively augments target recognition and aesthetic quality under relatively straightforward image conditions. In contrast, on the more challenging TAD66K dataset, which features natural landscapes and multi-element scenes, both models demonstrate even more pronounced gains, indicating that, amid higher scene complexity and aesthetic demands, AesX provides more substantial benefits for both diffusion and autoregressive paradigms, simultaneously bolstering scene parsing accuracy and the refinement of aesthetic details.

Attribute	From humans	By AesX
TIC	0.924	0.873
TQ	0.862	0.813
CF	0.824	0.818
CE	0.783	0.735
ER	0.792	0.765
AC	0.753	0.724

Table 5: SRCC results between MOSs and image attribute scores from humans and AesX, respectively.

From Table 3, we observe that both the autoregressive paradigm, represented by Emu3, and the diffusion paradigm, represented by DALLE3, exhibit relatively stable performance gains across multiple mainstream text-to-image datasets upon introduction of the AesX, demonstrating its strong generalization capacity. More specifically, taking Emu3 as an example: on GenEval, its overall score increases from 0.66 to 0.68, while the Color, Shape, and Texture metrics on T2I-CompBench improve from 0.7913, 0.5846, and 0.7422 to 0.8125, 0.6043, and 0.7601, respectively. Meanwhile, on DPG-Bench, the average score rises from 80.6 to 82.3, indicating that with AesX, the model achieves more refined control over color, shape, and texture details and can better fulfill aesthetic and generation requirements in complex scenarios.

5.4 Ablation Study

From Table 4, we observe that removing the knowledge enhancement module exerts a comparatively larger impact, indicating that knowledge enhancement plays a pivotal role in augmenting target recognition and scene detail. Excluding the preference enhancement module reduces the overall score, highlighting the critical contribution of preference enhancement in balancing multi-dimensional aesthetic preferences and textual alignment. A more granular perspective shows that omitting image-level preference enhancement (w/o image-level preference enhancement) and token-level preference enhancement (w/o token-level preference enhancement) causes progressively smaller drops, yet both still provide additional gains through multi-round iteration and

Figure 3: Effect of AesX in the progression from (a) to (d) in text-image consistency and image artistic score.

localized fine-tuning.

From Table 5, we observe that across six metrics, the Spearman rank correlation coefficients (SRCC) derived from the AesX closely align with those obtained from human evaluations, with overall discrepancies typically remaining below 0.1. In particular, on composition form, the two SRCC values are nearly identical, indicating that the image quality assessments conducted through the reward mechanism in AesX exhibit a high degree of congruence with human aesthetic judgments. This observation underscores AesX's reliability in evaluating aesthetic attributes and text alignment.

We further analyze the sensitivity of α and ϵ. Smaller α causes premature termination and insufficient refinement. As α increases, quality improves but quickly saturates, motivating our choice of α=5. For ϵ, we adopt a sample-relative stopping threshold rather than a fixed absolute value, since different samples exhibit substantially different score-change scales. A stricter ϵ risks missing later valid refinements, while a looser ϵ induces unnecessary deeper search without consistent gains.

5.5 Case Study

We conduct a case study to validate the effect of AesX in Fig. 3. The Fig. 3(a) column exhibits deficiencies in text-image consistency and aesthetic quality, whereas the Fig. 3(d) column demonstrates significant improvements in object accuracy, completeness, and overall aesthetic expression. In the first row, the absence of a person in Fig. 3(a) results in a mismatch between the textual description and the generated image, whereas Fig. 3(d) clearly presents the person, kites, beach, and ocean, enhancing text-image alignment, aesthetic quality, emotional resonance, and cultural significance. In the second row, although the teddy bear in Fig. 3(a) aligns with the textual description, the scene lacks emotional depth, while Fig. 3(d) employs warm lighting and a more harmonious composition to evoke a stronger emotional connection. Similarly, in the third row, the zebras in Fig. 3(a) appear relatively simplistic in detail, whereas Fig. 3(d) introduces richer colors and enhanced depth, creating a more visually appealing and artistically refined representation. Consequently, Fig. 3(d) outperforms Fig. 3(a) across all three rows in terms of text-image alignment, aesthetic quality, emotional resonance, and cultural significance.

6 Conclusions

Image aesthetic quality, which is critical for meeting human preferences, has immense practical value across diverse industries such as advertising, publishing, and digital arts. Despite the capacity of existing models to produce images with relatively high clarity and text alignment, finer aesthetic demands and the controllability of image quality remain unresolved challenges. Through our defined "Text-image consistency" and "Image Attribute Quality" metrics, as well as AesX, the quality of generated images by various models gains improvements in key aspects such as technical quality, composition form, etc. Experiments demonstrate that AesX achieves better performance in both target recognition accuracy and overall visual aesthetic presentation.

Acknowledgements

This work is supported by the Key-Area Research and Development Program of Guangdong Province (2024B0101050005), the Natural Science Foundation of Jiangsu Province (BK20251823), Beijing Natural Science Foundation (4262049), Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies (SZS2023007), Smart Social Governance Technology and Innovative Application Platform (YZCXPT2023101), the Innovation System of the Integration between Industry and Education for Smart Governance (CJRH2024101), and the Leadership Talent Program (Science and Education) of SIP.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Hangwei Chen, Feng Shao, Baoyang Mu, and Qiuping Jiang. 2024. Image aesthetics assessment with emotion-aware multibranch network. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. 2023. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yuyan Chen, Yifan Jiang, Li Zhou, Jinghan Cao, Yu Guan, Ming Yang, and Qingpei Guo. 2025. Engage for all: Making ordinary image descriptions appealing again! In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19342–19352.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis, 2024. *URL https://arxiv. org/abs/2403.03206*, 2.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686.

Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. 2024. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4744–4753.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. 2022. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, pages 942–948.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. 2024a. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5911–5920.

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

Minrui Jia, Guangao Wang, Zibei Wang, Shuai Yang, Yongzhen Ke, and Kai Wang. 2025. Self-supervised image aesthetic assessment based on transformer. *International Journal of Computational Intelligence and Applications*, 24(01):2450029.

Shuvue Jia, Baoliang Chen, Dingquan Li, and Shiqi Wang. 2022. No-reference image quality assessment via non-local dependency modeling. In *2022 IEEE 24th International workshop on multimedia signal processing (MMSP)*, pages 01–06. IEEE.

Yanyu Li, Xian Liu, Anil Kag, Ju Hu, Yerlan Idelbayev, Dhritiman Sagar, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. 2024. Textcraftor: Your text encoder can be image quality controller. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7985–7995.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411.

Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang. 2024. Text-to-image generation for abstract concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3360–3368.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. 2024. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. 2024. Dynamic prompt optimizing for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26627–26636.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. 2024. Diffusiongpt: Llm-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Christoph Redies, Ralf Bartho, Lisa Koßmann, Branka Spehar, Ronald Hübner, Johan Wagemans, and Gregor U Hayn-Leichsenring. 2024. A toolbox for calculating objective image properties in aesthetics research. *arXiv preprint arXiv:2408.10616*.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8552.

Derya Soydaner and Johan Wagemans. 2024. Multi-task convolutional neural network for image aesthetic assessment. *Ieee Access*, 12:4716–4729.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. 2024a. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024b. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024c. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Luwei Xiao, Xingjiao Wu, Junjie Xu, Weijie Li, Cheng Jin, and Liang He. 2024. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Information Fusion*, 106:102304.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Xingao Yan, Feng Shao, Hangwei Chen, and Qiuping Jiang. 2024. Hybrid cnn-transformer based meta-learning approach for personalized image aesthetics assessment. *Journal of Visual Communication and Image Representation*, 98:104044.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024a. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, and Tong Sun. 2024b. Customization assistant for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9182–9191.

A Quality Validation of AesBench

We first adopt GPT-4o for self-validation to verify whether the textual descriptions fully and accurately capture all objects in each image, scoring completeness and accuracy as binary indicators (0 or 1). Any image whose score is not 1 in either metric is regenerated until both scores reach 1. Next, 30% of the data is sampled for manual inspection by three graduate students; the completeness and accuracy scores must again both be 1. If an image fails to meet these requirements, the feedback is returned to GPT-4o for modification, and the remaining 70% is revalidated until all scores are 1. A different 30% subset is then sampled for a second round of manual inspection, and the procedure is repeated until the manual validation scores are 1 for all images with a 100% agreement among the three annotators, ultimately yielding 5,000 entries.

B Details of Datasets and Baselines

The selected datasets contain image aesthetic datasets, such as SPAQ (Fang et al., 2020) and TAD66K (He et al., 2022), as well as text-to-image datasets, such as MSCOCO (Chen et al., 2015), GenEval (Ghosh et al., 2023), T2I-CompBench (Huang et al., 2023) and DPG-Bench (Hu et al., 2024).

Specifically, SPAQ comprises 11,125 real-world images taken with 66 different smartphones, accompanied by EXIF metadata that facilitate further analysis of scene brightness and camera settings. TAD66K is a heavily annotated, thematically diverse image dataset with targeted evaluation criteria, effectively mitigating long-tailed distributions. MSCOCO 30K focuses on natural images containing multiple objects and associated captions, making it ideal for assessing a model's ability to handle complex scenarios. GenEval deconstructs textual prompts into discrete objects, attributes, and relations, thereby evaluating the model's capacity to satisfy requirements at different compositional levels. T2I-CompBench addresses multi-level compositional skills, such as color, shape, and texture binding, and underscores the model's robustness in synthesizing multiple concepts. DPG-Bench supplies 1,065 prompts containing numerous objects and relationships in extended text descriptions, testing semantic alignment and detail fidelity in complex scenes.

The selected baselines contain autoregressive models and diffusion models. Autoregressive models include Emu (Sun et al., 2023), Show-o (Xie et al., 2024), Transfusion (Zhou et al., 2024a), Chameleon (Team, 2024), LlamaGen (Sun et al., 2024), and Emu3 (Wang et al., 2024b), which typically maintain tight text-image alignment via sequential generation strategies. Diffusion-based approaches encompass SDv1.5 (Ramesh et al., 2022), SDv2.1 (Ramesh et al., 2022), DALLE2 (Ramesh et al., 2022), SDXL (Podell et al., 2023), PixArt-alpha (Chen et al., 2023), DALLE3 (Betker et al., 2023), and SD3 (Esser et al., 2024), refining images progressively through the diffusion process to achieve high-resolution outputs. Leveraging this diverse range of baselines enables a thorough assessment of AesX's adaptability and performance under different generation paradigms.

Specifically, SDv1.5 and SDv2.1 are based on latent diffusion models, proposing a method for image synthesis in latent space by executing the diffusion process within a compressed, low-dimensional representation. DALLE2 adopts a hierarchical text-conditioned generation strategy, initially employing a diffusion prior to generate CLIP latents from textual input and subsequently decoding these embeddings into images via an image generation module. SDXL improves upon the stable diffusion framework by optimizing both model architecture and training procedures, thereby enhancing the detail reproduction and complex scene restoration capabilities in high-resolution image synthesis. PixArt-alpha introduces a fast training methodology for diffusion transformers, capitalizing on the transformer's ability to capture long-range dependencies and fine structural details. DALLE3 builds upon its predecessor by refining the processing of captions and descriptions, leading to image outputs that better correspond to the intended semantic expression and detail presentation. SD3 scales the diffusion generation process by employing scaling rectified flow transformers that strike a balance between detailed recovery and overall compositional coherence in high-resolution images. Emu focuses on generative pretraining in a multimodal context, concurrently learning joint representations for both text and images to bridge generation and understanding tasks across modalities. Show-o proposes a unified transformer approach for multimodal understanding and generation, enabling efficient interaction and fusion of image and textual information within a single model. Transfusion integrates next-token prediction with the image diffusion process

within one multimodal framework. Chameleon employs an early-fusion strategy for mixed modalities, integrating disparate modal information at the initial stages of processing. LlamaGen challenges conventional diffusion paradigms by adopting an autoregressive approach based on the Llama architecture. Emu3 further advances this concept by asserting that next-token prediction alone suffices, contending that with adequate data and computational resources, autoregressive modeling is capable of capturing the complex distribution inherent in generated images.

For machine metrics, zero-shot FID (Fréchet Inception Distance) is computed following Wang et al. (2024b) by randomly selecting 30,000 text prompts from the validation set. We measure CLIP-T and CLIP-I scores based on CLIP-ViT-B or CLIP-ViT-L to assess text adherence and image quality. Classifier-free guidance (Ho and Salimans, 2022) is employed for image generation, with a default guidance scale of 5.0 unless stated otherwise. For the MSCOCO 30K dataset, performance results for other models come from Sun et al. (2023); Wang et al. (2024b); Xie et al. (2024); Zhou et al. (2024a). On GenEval, covering six dimensions (Single Object, Two Objects, Counting, Colors, Position, and Color Attribute), we evaluate in line with Esser et al.; Ghosh et al. (2023); Wang et al. (2024b); Xie et al. (2024); Zhou et al. (2024a), report outcomes using GPT-4V as a rewriter. For T2I-CompBench, we adopt a BLIP-VQA model to evaluate color, shape, and texture, referencing (Betker et al., 2023; Chen et al., 2023; Feng et al., 2024). On DPG-Bench, we rely on the mPLUG-large model to assess generated images according to the methods in Hu et al. (2024); Liu et al. (2024).

C Limitations

AesX currently relies on GPT-4o for prompt expansion, data validation, and reward modeling. This introduces concerns around reproducibility and potential aesthetic bias, as GPT-4o's internal aesthetic logic is not fully transparent. To mitigate circularity, human annotation and manual validation are incorporated in AesBench construction (Table 5). Future work may replace GPT-4o with open-source alternatives and conduct cross-model validation. Additionally, abstract metrics such as cultural connotation may not fully capture demographic-specific nuances, and AesX should be understood as one possible optimization scheme rather than a universal aesthetic standard.