Why Settle for One?
Text-to-ImageSet Generation and Evaluation

Anonymous Author(s)
Affiliation
Address
email

Figure 1: Illustration of diverse applications of Text-to-ImageSet generation across multiple domains. All examples are from the proposed benchmark. Images in the upper are generated by our AutoT2IS, while images in the lower are generated by AutoT2IS combined with the commercial model GPT-4o.

Abstract

Despite remarkable progress in Text-to-Image models, many real-world applications require generating coherent image sets with diverse consistency requirements. Existing consistent methods often focus on a specific domain with specific aspects of consistency, which significantly constrains their generalizability to broader applications. In this paper, we propose a more challenging problem, Text-to-ImageSet (T2IS) generation, which aims to generate sets of images that meet various consistency requirements based on user instructions. To systematically study this problem, we first introduce **T2IS-Bench** with 596 diverse instructions across 26 subcategories, providing comprehensive coverage for T2IS generation. Building on this, we propose **T2IS-Eval**, an evaluation framework that transforms user instructions into multifaceted assessment criteria and employs effective evaluators to adaptively assess consistency fulfillment between criteria and generated sets. Subsequently, we propose **AutoT2IS**, a training-free framework that maximally leverages pretrained Diffusion Transformers' in-context capabilities to harmonize visual elements to satisfy both image-level prompt alignment and set-level visual consistency. Extensive experiments on T2IS-Bench reveal that diverse consistency challenges all existing methods, while our AutoT2IS significantly outperforms current generalized and even specialized approaches. Our method also demonstrates the ability to enable numerous underexplored real-world applications, confirming its substantial practical value. All our data and code will be publicly available.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1 Introduction

Recent advancements in Text-to-Image (T2I) models [3, 4, 11, 24, 31, 37, 39] have substantially enhanced the ability to generate visually compelling and semantically faithful images. However, as illustrated in Figure 1, most practical scenarios, *e.g.*, product design, process illustrations, or character creation, often require not a single image, but a coherent image set. Such image set must maintain ***set-level visual consistency*** with varying degrees of ***identity preservation*** [19], ***uniform style*** [16], and ***logical coherence*** [41]. This novel requirement poses huge challenges for current T2I models, which are primarily focused on image-level prompt alignment rather than set-level visual consistency.

Previous attempts addressing this coherence have largely concentrated on specific facets of consistency [25, 41, 48] or specific combinations of consistency types [5, 30]. For instance, consistent character generation approaches [25, 29, 45] aim to preserve character identity across various contexts, while methods such as StoryDiffusion [51] are tailored for generating stylistically coherent comic sequences with localized identity preservation. Despite promising progress within specific domains, they heavily rely on specialized data, techniques, and model architectures, suffering from additional resource and model design costs. More importantly, their focus on such specific aspects of visual coherence significantly limits their generalizability to a wider variety of applications. To this end, we propose a more challenging problem, ***Text-to-ImageSet (T2IS)*** generation, which aims to receive diverse user instructions and generate image sets that satisfy multifaceted consistency requirements. We address this challenge through two key contributions:

(1) T2IS Benchmark: We first present **T2IS-Bench**, the pioneering benchmark crafted specifically for T2IS generation. Derived from a comprehensive collection of real-world requirements, T2IS-Bench encompasses 596 representative tasks distributed across 26 detailed subcategories. Figure 1 showcases illustrative examples of diverse tasks. Based on this, we introduce **T2IS-Eval**, an evaluation framework that automatically transforms user instructions into multifaceted assessment criteria across three aspects of consistency: ***identity***, ***style***, and ***logic***. Our framework then leverages the strong multi-image recognition capabilities of large-scale models [2] to serve as effective consistency evaluators and obtain logits of "Yes-or-No" answers for each criterion to assess the consistency fulfillment. More importantly, unlike previous methods that focus on a single aspect of consistency [14, 17], T2IS-Eval leverages transformed criteria to enable adaptive and interpretable consistency evaluation.

(2) T2IS Generation: We propose **AutoT2IS**, a training-free framework that maximally leverages the remarkable in-context generation capabilities of pretrained Diffusion Transformers (DiTs) [21, 24]. Specifically, AutoT2IS first employs a structured recaptioning approach to systematically parse user instructions into informative prompts for both individual images and the global image set. Subsequently, we introduce a novel set-aware generation with the divide-and-conquer strategy: first binding individual images' prompts to their respective independent latents during early denoising stages to establish unique content characteristics, then integrating these latents through the multi-modal attention mechanism with a global consistency prompt. This enables each visual latent to simultaneously attend to its individual prompt, the global consistency prompt, and other images' visual latents. This way dynamically harmonizes visual elements to satisfy both image-level prompt alignment and set-level visual consistency, effectively bridging the gap between isolated image generation and set-level consistency without requiring specialized fine-tuning.

We conduct comprehensive experiments on the proposed T2IS-Bench to evaluate various approaches, including unified models (*e.g.*, Show-o [47], Janus-Pro [7]), compositional frameworks (*e.g.*, Gemini+Flux [24]), and agentic frameworks (*e.g.*, ISG-Agent [5], ChatDiT [22]). Our evaluation results indicate that while current methods have achieved excellent performance on prompt alignment, there remains a significant gap in achieving set-level consistency, highlighting the value of this underexplored research direction. Furthermore, both quantitative and qualitative results demonstrate that our proposed AutoT2IS substantially outperforms existing approaches, achieving significant improvements across three aspects of consistency. Notably, AutoT2IS also achieves competitive or even superior performance in specific domains compared to specialized methods.

Interestingly, during the development of this work, advanced commercial image generation models like Gemini 2.0 [15] and GPT-4o [31] were released. This positions T2IS-Bench as the first platform to systematically evaluate and analyze their T2IS capabilities. Our assessments reveal that these models excel in consistency but often sacrifice image quality and prompt alignment. To this end, as shown in the bottom of Figure 1, AutoT2IS can be seamlessly integrated with models like GPT-4o

to further boost performance. By offering a transparent benchmark alongside a robust baseline, our work lays a strong foundation for the advancement of T2IS research.

2 Related Work

Text-to-Image Generation. With the advancement of diffusion models [18, 40], Text-to-Image (T2I) models [3, 4, 11, 12, 24, 31, 33, 35–37, 39] have demonstrated exceptional capabilities in generating high-quality images that accurately align with textual descriptions. Early approaches, such as DALL-E 2 [36] and Stable Diffusion [33], employed Latent Diffusion Models [37] (LDMs) with U-Net backbones [38] for efficient denoising in compressed latent space. Recent developments have advanced the field through the integration of transformer architectures into diffusion models, known as Diffusion Transformers (DiTs) [12, 24, 32]. These models leverage global attention mechanisms to capture complex dependencies in data, resulting in improved scalability and image fidelity. However, existing methods primarily focus on generating individual images, with limited exploration into the generation of coherent image sets. In this paper, we investigate and harness the in-context generation capabilities of DiTs to maximize their potential for addressing diverse T2IS tasks.

Consistent Text-to-Image Generation. Existing methods related to T2IS generation span various domains, primarily including consistent character generation [25, 29, 45] and storytelling [30, 51]. In character generation, training-based approaches such as PhotoMaker [25] aim to preserve identity via parameter-efficient fine-tuning, while training-free methods like One-Prompt-One-Story [29] exploit the contextual consistency of language models without additional training. In contrast to character generation, storytelling further demands style coherence across images. Make-a-Story [34] incorporates a visual memory module to capture contextual cues, and StoryDiffusion [51] introduces Consistent Self-Attention to enhance both identity and style consistency. However, when tackling new tasks like process generation [41] that require strong logical consistency for generating drawing processes, previous methods often fall short and need customized task-specific fine-tuning [41]. This suggests that consistencies in existing methods are generally domain-specific and lack generalizability. In this work, we propose T2IS generation to address a broad range of consistency requirements.

3 Methodology

Our goal is to develop T2IS methods that aim to generate coherent image sets from diverse user instructions. To achieve this objective, we first provide a comprehensive T2IS benchmark with an evaluation framework for diverse tasks in Sec. 3.1. In Sec. 3.2, we introduce our generation method, AutoT2IS, which consists of Structured Recaption and Set-Aware Generation to generate image sets that satisfy both image-level prompt alignment and set-level visual consistency.

3.1 T2IS Benchmark

T2IS-Bench. We first construct a comprehensive benchmark that reflects real-world demands on diverse consistency, as illustrated in Figure 2. The construction follows a multi-stage process:

First, we conduct an extensive analysis of real-world T2IS applications and categorize these into five major groups, covering diverse consistency requirements. For instance, Character Generation primarily focuses on identity preservation, while Process Generation emphasizes logical consistency. It is noteworthy that falling within the same groups does not mean the requriements of consistency are exactly the same. Rather, categories within each group vary in the extent to which they integrate and balance additional dimensions of consistency.

Second, we collect diverse data from two complementary sources. We incorporate established benchmarks including Consistency+ [29], ISG-Bench [5], IDEA-Bench [27] and others [21, 23, 41]. Additionally, we curate examples from real-world platforms, such as e-commerce websites (*e.g.*, Amazon, Taobao), image generation communities (*e.g.*, Civitai), social media platforms (*e.g.*, RedNote), and other sources. From academic benchmarks, we select representative examples based on their diversity, while from platforms, we prioritize examples with user engagement and popularity.

Third, we standardize all these examples by employing MLLMs such as GPT-4o to transform these into conversational user instructions. We further augment the dataset through in-context learning techniques [10] to ensure balanced samples across all subcategories. All samples undergo rigorous

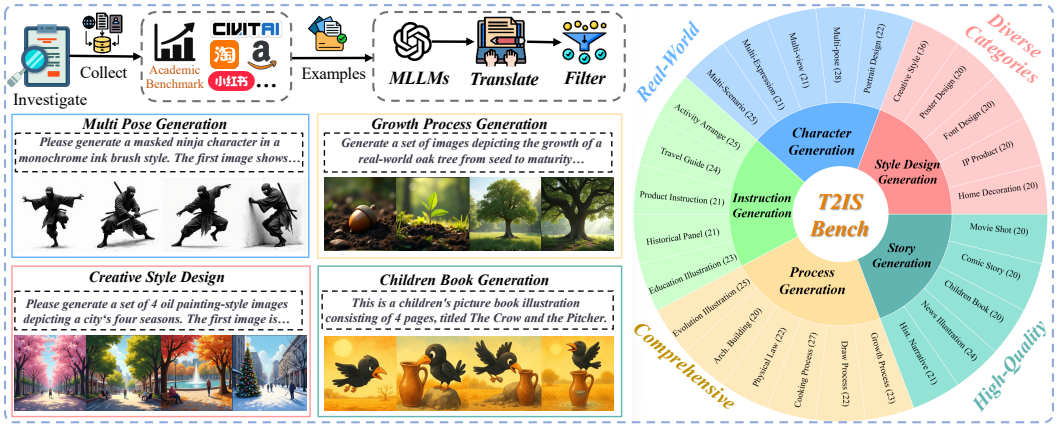

Figure 2: Illustration of the proposed T2IS-Bench. Upper Left: The collection process. Lower Left: Task examples displays. Right: Categories with their corresponding quantities.

verification by PhD-level domain experts, with duplicate or highly similar instances eliminated through a combination of automated similarity detection using BertScore [50] and manual review.

Through the above process, we compile **T2IS-Bench**, containing 596 high-quality user instructions distributed across 26 distinct subcategories. Our benchmark offers the most comprehensive coverage to date, establishing a solid foundation for advancing T2IS. Detailed information regarding category definitions, representative examples, and dataset distribution statistics is provided in Appendix A.

T2IS-Eval. We then propose **T2IS-Eval**, a comprehensive evaluation framework that assesses image sets across three critical dimensions: ***identity***, ***style***, and ***logical consistency***. For each dimension, we employ LLMs to generate targeted evaluation questions as assessment criteria that reflect the consistency requirements corresponding to the current user instructions. As illustrated in Figure 3, the criteria across different dimensions include:

- **Identity** criteria measure the consistency of identity preservation across images. This includes whether specific entities (*e.g.*, characters, objects) maintain consistent appearance features (*e.g.*, size, color, shape), and whether their emotional expressions or key identifying elements persist.
- **Style** criteria assess the style uniform across images, including consistency in illustration style (*e.g.*, watercolor, cartoon, 3D rendering), harmony in color palette usage, and others.
- **Logical** criteria evaluate whether the image set maintains reasonable causal relationships and narrative coherence. This includes environmental consistency across scenes, proper depiction of cause-and-effect relationships, and logical alignment between actions and their consequences.

Then, we systematically assess how well image sets satisfy these criteria. Inspired by VQAscore [28], we discovered that large-scale MLLM [2] with strong multi-image recognition capabilities can serve as excellent consistency judges. Based on this insight, we formulate our evaluation criteria as questions that are sequentially applied to image pairs, obtaining logit scores for "Yes-or-No" answers that quantify the degree of consistency fulfillment. In Appendix B, we demonstrate the effectiveness of this approach and its advantages over direct scoring by MLLMs. Overall, unlike previous methods that focus on a single consistency dimension [14], our approach leverages transformed criteria to enables adaptive consistency evaluation with fine-grained measurements across multiple dimensions.

In addition to consistency evaluation, we also incorporate established methods to comprehensively assess generation quality. Specifically, we employ MPS [49] to evaluate the aesthetic quality of generated images, and VQAscore [28] to measure the alignment between each image and its corresponding instruction. Notably, all three evaluation components (aesthetics, prompt alignment, and visual consistency) leverage vision-language models and employ a unified scoring mechanism based on next-token prediction logits or text-image matching logits. Consequently, these quantitative metrics are normalized to a consistent range $[0, 1]$, providing better interpretability and holistic assessment.

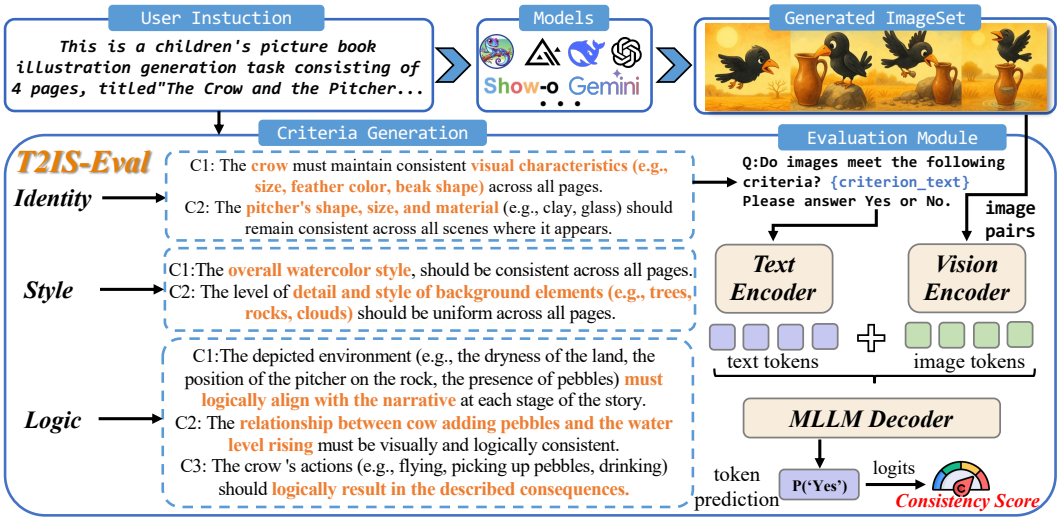

Figure 3: Overview of our T2IS-Eval. For each dimension, LLMs are prompted to generate 2-4 criteria; we only display a subset. Highlighted portions show the core elements for evaluation.

3.2 T2IS Generation

In this section, we introduce AutoT2IS, a training-free framework for T2IS generation. We aim to maximally leverage the remarkable in-context generation capabilities [21] of pretrained Diffusion Transformers (DiTs [24, 32]) to handle diverse T2IS tasks. We delineate two key steps, as depicted in Figure 4. Specifically, AutoT2IS first utilizes a ***structured recaption*** to parse user instructions into informative prompts for both individual images and the complete image set, thereby capturing richer semantic information. Subsequently, we introduce ***set-aware generation*** that effectively applies the recaptioned individual prompts and consistency requirements to the entire image set, transforming semantic contents into visually consistent outputs.

Structured Recaption. Let y denote the given user instruction, which contains requirements for individual images and overall consistency. We first employ LLMs to identify key content for each image and consistency requirements for the entire set. The recognized elements are represented in a structured format $S = \{E, C\}$, where $E = \{e_1, e_2, ..., e_n\}$ represents the content for n individual images, and $C = \{c_1, c_2, ..., c_m\}$ denotes the m consistency requirments extracted from y.

Based on the structured information S and original user instruction y, we generate enhanced prompts for each image and global consistency requirements:

$$p_i = f_{\text{recap}}(e_i, y)\ \forall i \in \{1, 2, ..., n\}, \quad g = f_{\text{consist}}(C, y) \tag{1}$$

where p_i represents the detailed prompt for the i-th image with denser fine-grained details to improve fidelity, and g represents the global consistency description that ensures the model adheres to overarching principles across all images. Functions f_{recap} and f_{consist} are implemented using LLMs to expand the structured information into comprehensive textual descriptions.

Set-Aware Generation. While DiT models exhibit powerful in-context capabilities, existing approaches [20, 21, 42] often require fine-tuning to activate these abilities, which compromises the model's original generalization performance. As shown in Figure 4, we introduce set-aware generation that enables consistency requirements to be expressed through textual prompts and multi-modal attention in a training-free way.

We empirically find that these few binding steps are sufficient to establish the unique content of each image, allowing subsequent steps to focus on resolving consistency issues across the image set. Inspired by this, we design a *divide-and-conquer* strategy to separate DiT's native denoising capabilities into two distinct phases. In the *divide* step, which is applied during the early stages of the denoising process with a total timestep t, we generate independent noise latents x_t^i for each image $i \in \{1, 2, ..., n\}$ and denoise them separately with their corresponding prompts p_i. This binding process is executed only during the initial r steps of the denoising process.

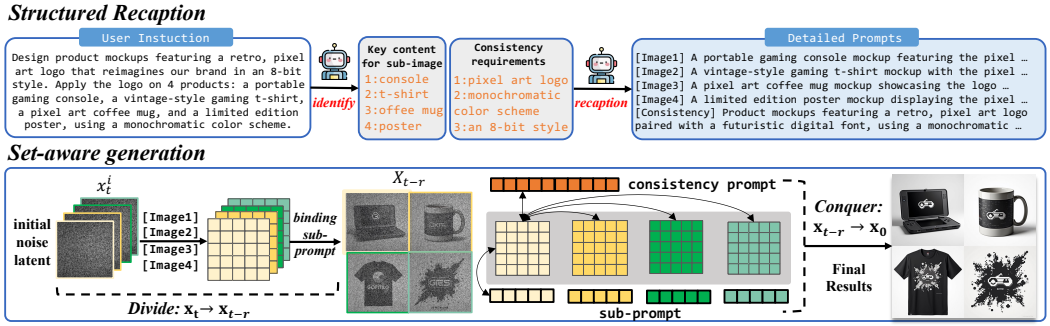

Figure 4: The pipeline of our AutoT2IS, consists of structured recaption and set-aware generation.

In the *conquer* phase, we concatenate all independently denoised latents x^i_{t-r} into a grid layout $X_{t-r} = [x^1_{t-r}, x^2_{t-r}, ..., x^n_{t-r}]$. This concatenation enables previously isolated regions to collaborate through DiT's multi-modal attention mechanism [12, 32]. Formally, for each denoising step $t' \in \{t-r, t-r-1, ..., 0\}$, the attention computation can be expressed as:

$$\text{Attention}(Q, K, V, M) = \text{softmax}\left(\frac{QK^T \odot M}{\sqrt{d_k}}\right)V \tag{2}$$

where $Q = W_Q X_{t'}$ represents queries derived from the concatenated latent $X_{t'}$, $K = [W^{\text{text}}_K p; W^{\text{text}}_K g; W^{\text{image}}_K X_{t'}]$ and $V = [W^{\text{text}}_V p; W^{\text{text}}_V g; W^{\text{image}}_V X_{t'}]$ index keys and values from both the prompt embeddings $P = [p_1, p_2, ..., p_n, g]$ and the latent representation itself. The mask matrix $M \in \mathbb{R}^{N \times (N_p + N_g + N)}$ is defined as:

$$M_{i,j} = \begin{cases} 1, & \text{if } i \in x^k_{t'} \text{ and } j \in \{p_k, g, X_{t'}\} \\ 0, & \text{otherwise} \end{cases}, \quad \text{where } k \in \{1, 2, \ldots, n\} \tag{3}$$

This mask ensures that each visual latent from a specific region $x^k_{t'}$ to simultaneously attend to three critical elements, as illustrated in Figure 4:

- Its corresponding individual prompt p_i for precise semantic content preservation;
- The global consistency prompt g for harmonious coherence across the image set;
- Visual latents in other images for enhanced cross-image alignment and contextual integration.

Through this enhanced multi-modal attention mechanism, the model dynamically harmonizes visual elements to satisfy both local alignment and global consistency requirements, effectively bridging the gap between isolated image generation and set-level consistency without specialized fine-tuning.

Extensions to image-conditioned and long-sequence generation. Our method extends naturally to image-conditioned and long-sequence T2IS generation. For image-conditioned tasks, we apply inversion techniques [13, 46] to extract latent tokens from reference images, then implement token replacement to preserve these conditioned regions while only denoising the remaining areas. This maintains visual consistency with reference images. For long-sequence generation (more than 5 images), we employ a sliding window approach that generates new images based on previously created windows, ensuring consistency across extended sequences while managing computational constraints. Additional details are provided in Appendix C.

4 Experiments

Implementation Details. We utilize DeepSeek-R1 [9] as the foundational LLM for criteria generation and structured recaptioning. Qwen2.5-VL-7B [2] serves as our consistency evaluator. For Set-Aware generation, we employ FLUX.1-dev [24] as the base model, executing a total of 20 denoising steps, i.e., 2 steps for the divide phase and the remaining for the conquer phase. For the concatenated latent, we apply positional encoding at new positions to preserve the original multi-modal attention structure.

Table 1: Comparisons with various generalized generation models on T2IS-Bench. The upper shows open-source models, while the lower presents commercial models. The best and second-best results among the open-source models are highlighted in bold and underlined, respectively.

Model	Aesthetics	Prompt Alignment			Visual Consistency			Avg.
		Entity	Attribute	Relation	Identity	Style	Logic	
Anole [8]	0.170	0.534	0.611	0.570	0.115	0.148	0.090	0.264
Show-o [47]	0.206	0.780	0.785	0.776	0.233	0.287	0.285	0.409
Janus-Pro [7]	0.333	0.787	0.785	0.781	0.224	0.272	0.300	0.435
Gemini & SD3 [12]	0.500	**0.796**	**0.794**	**0.789**	0.287	0.244	0.320	0.480
Gemini & Pixart [6]	0.447	0.743	0.765	0.747	0.206	0.279	0.268	0.440
Gemini & Hunyuan [26]	0.410	0.758	0.774	0.765	0.197	0.276	0.271	0.436
Gemini & Flux-1 [24]	**0.533**	0.791	0.790	0.786	0.249	0.302	0.328	0.490
ChatDit [22]	0.414	0.717	0.726	0.726	0.296	0.326	0.310	0.455
ISG-Agent [5]	0.256	0.667	0.703	0.682	0.146	0.178	0.163	0.338
AutoT2IS (ours)	0.520	0.729	0.756	0.743	**0.359**	**0.414**	**0.356**	**0.515**
Gemini 2.0 Flash [15]	0.430	0.738	0.747	0.743	0.428	0.383	0.392	0.509
GPT-4o [31]	0.445	0.663	0.683	0.693	0.400	0.463	0.383	0.501
AutoT2IS + GPT-4o	0.567	0.754	0.761	0.763	0.441	0.520	0.416	0.571

Due to space limitations, we present the most significant experimental results in the main text, including quantitative comparison of existing generalization and specialization methods with ours on T2IS-Bench (Sec. 4.1), qualitative visual comparisons to demonstrate the broad applications of our method (Sec. 4.2), and ablation studies on core components (Sec. 4.3). Please refer to the appendix for more comprehensive information. We provide detailed experimental settings in App.D, reliability analysis of T2IS-Eval in App.E, failure cases in App.G, limitations and future work in App.F.

4.1 Benchmarking T2IS: Quantitative Comparison of Generalized and Specialized Methods

Experiment Setups. We evaluate diverse frameworks for T2IS generation capabilities: (1) unified models like Show-o; (2) compositional models that use Gemini [43] to generate sub-captions with state-of-the-art T2I models; and (3) agentic models such as ChatDit that incorporate multi-step planning and generation processes. We also evaluate commercial models including Gemini 2.0 Flash and GPT-4o. Detailed setups for implementing these models are provided in Appendix H.

Visual Consistency Challenges All Generalized Methods. As illustrated in Table 1, most models perform well on prompt alignment, with average scores all exceeding 0.5, demonstrating the significant progress T2I models have made in alignment capabilities. However, all models exhibit significant deficiencies in visual consistency. All open-source methods score below 0.35 across all consistency dimensions, highlighting the value of exploring this research direction. Comparatively, our method achieves the best consistency capabilities among open frameworks, showing notable improvements over previous methods and delivering the best overall performance.

Commercial Models Still Need Improvement. As shown in the bottom of Table 1, we evaluate state-of-the-art commercial models [15, 31]. These models, benefiting from large-scale pretraining, demonstrate superior instruction understanding and better consistency than open-source alternatives. However, our results reveal that they still struggle with T2IS generation tasks. During T2IS generation, their aesthetic quality and prompt alignment significantly deteriorate, even performing worse than many open-source models. This shortcoming can be mitigated by incorporating Auto-T2IS's structured reception to enrich semantic information, which achieves significant improvements in quality, alignment, and consistency. These findings underscore the necessity for improved T2IS task instruction comprehension in commercial models.

AutoT2IS Matches/Exceeds Specialized Methods. Table 2 provides comparisons between our unified generation approach and specialized methods. These specialized methods are typically tailored and optimized for specific domains, allowing them to perform exceptionally well in those areas but limiting their adaptability across a wide range of T2IS tasks. The results in their respective domains on T2IS show that our method achieves competitive or even superior performance in areas where specialized methods have been pre-trained. This suggests that AutoT2IS effectively explores

Table 2: Comparisons of specialized methods and our unified workflow across different T2IS domains.

Model	Aesthetics	A Prompt Alignment			Visual Consistency			Avg.
		Entity	Attribute	Relation	Identity	Style	Logic	
Character Generation: Multi-view, Multi-scenario, Portrait Design								
PhotoMaker [25]	0.413	0.793	0.799	0.762	0.482	0.458	0.450	0.552
OnePrompt [29]	0.600	0.805	0.807	0.766	0.520	0.512	0.470	0.590
X-Flux [1]	0.428	0.834	0.804	0.780	0.531	0.529	0.521	0.553
AutoT2IS (ours)	0.557	0.798	0.788	0.764	0.609	0.520	0.509	**0.619**
Style Design Generation: Creative Style, Font, IP product design								
IPAdapter-Flux [44]	0.514	0.856	0.833	0.846	0.386	0.455	0.517	**0.582**
AutoT2IS (ours)	0.475	0.789	0.791	0.792	0.339	0.391	0.478	0.533
Story Generation: Movie Shot, Comic Story, Children Book								
ICLora-Story [21]	0.408	0.711	0.739	0.702	0.078	0.190	0.119	0.361
StoryDiffusion [51]	0.581	0.684	0.728	0.670	0.182	0.207	0.141	0.413
Story-Adapter [30]	0.578	0.749	0.766	0.734	0.181	0.328	0.226	**0.463**
AutoT2IS (ours)	0.534	0.629	0.698	0.660	0.262	0.405	0.237	0.456
Process Generation: Growth, Draw, Building								
MakeAnything [41]	0.358	0.636	0.665	0.652	0.329	0.340	0.222	0.408
AutoT2IS (ours)	0.578	0.727	0.754	0.739	0.314	0.346	0.274	**0.493**

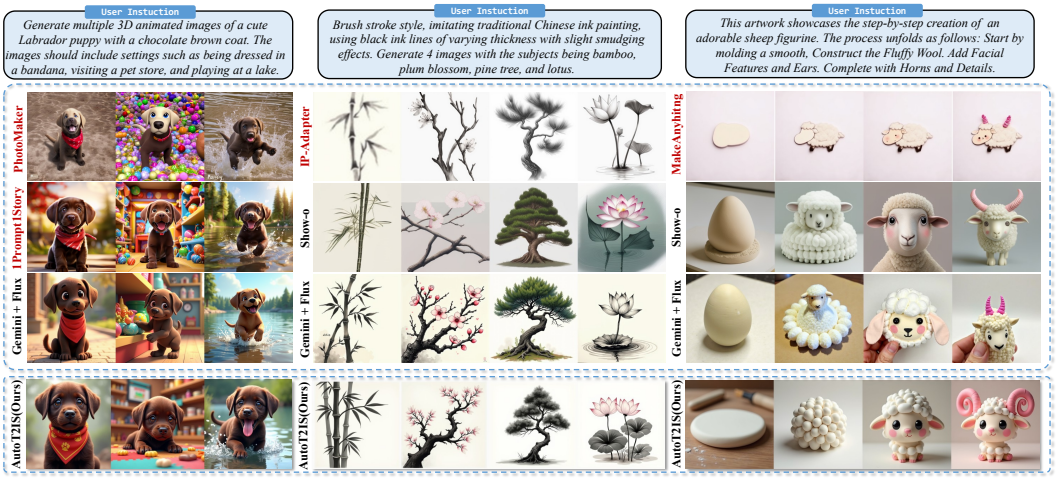

Figure 5: Qualitative comparison between our AutoT2IS and existing methods on common T2IS tasks. Black labels indicate generalized methods, while red labels represent specialized methods.

broader consistency principles while avoiding overfitting, thereby maintaining the model's inherent generative capabilities. These findings highlight the value of unified methods for diverse T2IS tasks.

4.2 Visualizing T2IS: Qualitative Comparison and Diverse Application Showcase

AutoT2IS Excels at Common T2IS Tasks. Figure 5 presents a qualitative comparison between our AutoT2IS and existing methods across several common T2IS tasks. As demonstrated, generalized methods like Show-o and Flux exhibit significant limitations in identity preservation, style uniform, and logical coherence. In contrast, AutoT2IS produces more visually consistent image sets. When evaluated against specialized methods in multi-scenario character generation, our approach maintains exceptional identity consistency while ensuring precise prompt alignment, achieving results comparable to the SOTA method 1Prompt1Story [29]. Notably, when it comes to challenging logical consistency generation tasks, existing methods like MakeAnything [41], despite being specifically trained for these scenarios, often compromise their fundamental generation capabilities, resulting in diminished image quality. Conversely, our training-free method preserves the model's original generation capabilities and simultaneously maximizing its inherent consistency potential.

Figure 6: AutoT2IS enables novel T2IS applications with significant real-world value.

AutoT2IS Enables More Real-world Applications. Beyond common T2IS tasks, our AutoT2IS enables to address previously underexplored applications with significant real-world value, as illustrated in Figure 6. These applications have been largely overlooked due to the lack of a unified approach to varying degrees of different consistency. For instance, IP Product Design (Task 2) requires coherence between IP concept identity and product style. Similarly, Growth Process Generation and Physical Law Illustration (Tasks 4 and 6) demand comprehensive consistency across identity, style, and logical dimensions. The diversity of these tasks demonstrates that focusing on a single aspect of consistency severely limits the range of potential applications, highlighting the value of unified methods for T2IS.

4.3 Ablation Studies

To better understand the contribution of each stage in AutoT2IS, we conduct ablation studies. Table 3 presents the quantitative results across different evaluation dimensions.

Importance of Structured Recaption (SR). The *structured recaption* plays a critical role by reformulating user inputs into more structured and detailed descriptions that better guide the image generation process. When we replace the *structured recaption* with a baseline approach that simply uses an LLM to generate sub-captions, we observe a significant drop across all evaluation metrics. These results highlight the importance of semantic completeness and relational coherence for effective T2IS generation.

Table 3: Ablation studies on key stages.

Variants	Aesthetics	Alignment	Consistency
Full	**0.520**	**0.742**	**0.376**
w/o SR	0.512	0.695	0.241
w/o SG	0.326	0.611	0.318

Importance of Set-Aware Generation (SG). The set-aware generation is responsible for creating a consistent set across multiple images, while ensuring the quality and alignment of each sub-image. We replace this stage with a baseline approach that directly concatenates prompts to guide the model in generating multi-grid images [21, 22]. As shown in results, despite achieving reasonable consistency, the quality of individual images and text-image alignment significantly decreased. This demonstrates the importance of the divide-and-conquer approach in Set-Aware Generation, where the model first understands the content of individual image before rendering them with visual consistency.

We include additional ablation with visualizations in Appendix I for more intuitive understanding.

5 Conclusion

In this paper, we advance the field of Text-to-ImageSet (T2IS) by introducing innovative frameworks for both T2IS generation and evaluation. We present T2IS-Bench, a comprehensive benchmark featuring 596 representative tasks across 26 detailed subcategories. Alongside this, we introduce T2IS-Eval, an evaluation framework that enables diverse consistency assessment. Furthermore, we introduce AutoT2IS, a training-free framework that fully exploits the exceptional in-context generation capabilities of DiT models, effectively harmonizing visual elements to meet both image-level prompt alignment and set-level visual consistency. By providing a transparent benchmark and a robust baseline, our work establishes a solid foundation for the future of T2IS research.

References

[1] XLabs AI. x-flux: Flux model fine-tuning scripts. https://github.com/XLabs-AI/x-flux, 2024. Apache License 2.0.

[2] 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. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Jason Baldridge, Kevin Shih, Yael Li, Zongyi Wang, Shweta Prabhudesai, Sashi Cheluvaraju, Xin Li, Lucy Chai, Miao Ding, Bryan Catanzaro, et al. Imagen 2: Tuning text-to-image diffusion models for photorealism and generalization. *arXiv preprint arXiv:2401.18680*, 2024.

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

[5] Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang, Yao Wan, Pan Zhou, and Ranjay Krishna. Interleaved scene graph for interleaved text-and-image generation assessment. *arXiv preprint arXiv:2411.17188*, 2024.

[6] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[8] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

[9] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[10] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[11] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[12] 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. In *Forty-first International Conference on Machine Learning*, 2024.

[13] Haoran Feng, Zehuan Huang, Lin Li, Hairong Lv, and Lu Sheng. Personalize anything for free with diffusion transformer. *arXiv preprint arXiv:2503.12590*, 2025.

[14] 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. In *Advances in Neural Information Processing Systems*, volume 36, pages 50742–50768, 2023.

[15] Google Cloud. Generate images with gemini | generative ai on vertex ai, 2025. URL https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/image-generation. Accessed: 2025-05-09.

[16] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024.

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[19] Jiehui Huang, Xiao Dong, Wenhui Song, Hanhui Li, Jun Zhou, Yuhao Cheng, Shutao Liao, Long Chen, Yiqiang Yan, Shengcai Liao, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024.

[20] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Huanzhang Dou, Yupeng Shi, Yutong Feng, Chen Liang, Yu Liu, and Jingren Zhou. Group diffusion transformers are unsupervised multitask learners. 2024.

[21] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.

[22] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Chen Liang, Tong Shen, Han Zhang, Huanzhang Dou, Yu Liu, and Jingren Zhou. Chatdit: A training-free baseline for task-agnostic free-form chatting with diffusion transformers. 2024.

[23] Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*, 2024.

[24] Black Forest Labs. Flux: Inference repository. `https://github.com/black-forest-labs/flux`, 2024. Accessed: 2024-10-25.

[25] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[26] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024.

[27] Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Junge Zhang, Xin Zhao, and Yu Liu. Idea-bench: How far are generative models from professional designing? *arXiv preprint arXiv:2412.11767*, 2024.

[28] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024.

[29] Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=cD1kl2QKv1`.

[30] Jiawei Mao, Xiaoke Huang, Yunfei Xie, Yuanqi Chang, Mude Hui, Bingjie Xu, and Yuyin Zhou. Story-Adapter: A Training-free Iterative Framework for Long Story Visualization, 2024.

[31] OpenAI. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025. Accessed: 2025-03-25.

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

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

[34] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2493–2502, 2023.

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Tim Salimans, Jonathan Ho, David J Fleet, Phillip Isola, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[41] Yiren Song, Cheng Liu, and Mike Zheng Shou. Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. *arXiv preprint arXiv:2502.01572*, 2025.

[42] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

[43] 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.

[44] InstantX Team. Instantx flux.1-dev ip-adapter page, 2024.

[45] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43 (4):52, 2024. doi: 10.1145/3658157. URL https://doi.org/10.1145/3658157.

[46] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

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

[48] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

[49] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024.

[50] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SkeHuCVFDr`.

[51] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly and accurately reflect our core contributions and scope. We introduces a new problem setting (Text-to-ImageSet generation), constructs a benchmark dataset (T2IS-Bench), proposes a corresponding evaluation framework (T2IS-Eval), and presents a training-free solution (AutoT2IS). The claims are supported by experimental results demonstrating the superiority of AutoT2IS over existing approaches and its generalization across a broad set of real-world applications.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The detailed limitation analyses are available in the Appendix F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not contain any theoretical results, assumptions, or formal proofs. Its contributions are empirical and methodological in nature, focusing on benchmark construction, evaluation framework design, and a training-free generative method.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The detailed experimental settings are available in the Experiments section 4 of the main paper and the Appendix D. All data and code will be made public.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the details of benchmark data in Appendix A, and detailed experimental settings in Appendix D. All data and code will be made public upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The detailed experimental settings are available in the Experiments section 4 of the main paper and the Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We conduct multiple runs to ensure the reliability of the evaluation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are performed using NVIDIA L40s GPUs. The detailed experimental settings are available in the Experiments section 4 of the main paper and the Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. All datasets used in this study are publicly available and widely used in the community. No personally identifiable information or sensitive content is involved.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper discusses both the positive and potential negative societal impacts. On the positive side, the proposed T2IS task and framework could significantly improve the controllability and practical utility of text-to-image generation models, enabling more coherent and useful image generation for applications such as education, creative design, e-commerce, and assistive tools. However, the work also raises potential concerns: generating consistent image sets based on user instructions could be misused to mass-produce coordinated synthetic images for misinformation, propaganda, or impersonation purposes. To mitigate these risks, we suggest responsible release of model checkpoints and emphasize the importance of watermarking, usage monitoring, and ethical deployment practices.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper acknowledges the potential risks of misuse associated with releasing image generation frameworks and datasets. To mitigate these concerns, the released T2IS-Bench dataset has undergone a manual filtering process to ensure that it does not contain unsafe, offensive, or privacy-violating content. For the AutoT2IS model code, we provide detailed usage documentation and include explicit terms of use to discourage malicious applications such as mass-produced disinformation. In addition, we do not release fine-tuned weights or specialized prompt templates that could be easily misused without further safeguards. We encourage responsible research use and welcome community feedback to further improve safety practices.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper makes use of publicly available pretrained Diffusion Transformers and other generative models, such as Stable Diffusion, which are properly cited in the paper. We explicitly mention the versions used and adhere to their respective licenses—for instance, Stable Diffusion under the CreativeML Open RAIL-M license. For any datasets used for evaluation or benchmarking, we ensure their original creators are credited and that their terms of use are respected. All reused assets are only employed within the allowed scope of their licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces several new assets, including the T2IS-Bench dataset, the T2IS-Eval evaluation toolkit, and the AutoT2IS framework. All assets are well documented and will be released with accompanying instructions on usage, data format, evaluation protocols, and limitations. The dataset construction process is described in detail in the paper, and any content included in T2IS-Bench was either synthetically generated or curated from publicly available and appropriately licensed sources, avoiding privacy violations or copyrighted material. At submission time, an anonymized version of the assets is provided to ensure compliance with double-blind review requirements. Upon publication, all assets will be released under a permissive license with appropriate citation and usage guidelines.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This work does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This work does not involve human subjects or crowdsourcing experiments and therefore does not require IRB or equivalent ethical approval.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Large language models (LLMs) are used as important and non-standard components in both the AutoT2IS generation framework and the T2IS-Eval evaluation pipeline. In AutoT2IS, LLMs are employed during the semantic captioning stage to generate high-level semantic descriptions of generated images, which are then used to guide visual consistency alignment. In T2IS-Eval, LLMs are used to parse diverse user instructions and convert them into structured, multi-faceted evaluation criteria for consistency assessment. These uses of LLMs play a central role in ensuring instruction compliance and semantic alignment throughout the generation and evaluation process. The methodology and potential limitations of LLM usage are discussed in detail in the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.