Rethinking The Training And Evaluation of Rich-Context Layout-to-Image Generation

Jiaxin Cheng[2*] **Zixu Zhao**[1] **Tong He**[1] **Tianjun Xiao**[1] **Zheng Zhang**[1] **Yicong Zhou**[2]
{yc47434,yicongzhou}@um.edu.mo {zhaozixu,tianjux,htong,zhaz}@amazon.com
[1]Amazon Web Services Shanghai AI Lab [2]University of Macau

Figure 1: The proposed method demonstrates the ability to accurately generate objects with complex descriptions in the correct locations while faithfully preserving the details specified in the text. In contrast, existing methods such as BoxDiff [57], R&B [56], GLIGEN [25], and InstDiff [54] struggle with the complex object descriptions, leading to errors in the generated objects.

Abstract

Recent advancements in generative models have significantly enhanced their capacity for image generation, enabling a wide range of applications such as image editing, completion and video editing. A specialized area within generative modeling is layout-to-image (L2I) generation, where predefined layouts of objects guide the generative process. In this study, we introduce a novel regional cross-attention module tailored to enrich layout-to-image generation. This module notably improves the representation of layout regions, particularly in scenarios where existing methods struggle with highly complex and detailed textual descriptions. Moreover, while current open-vocabulary L2I methods are trained in an open-set setting, their evaluations often occur in closed-set environments. To bridge this gap, we propose two metrics to assess L2I performance in open-vocabulary scenarios. Additionally, we conduct a comprehensive user study to validate the consistency of these metrics with human preferences. https://github.com/cplusx/rich_context_L2I/tree/main

*Work done during his internship at Amazon

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1983

1 Introduction

Recent years have witnessed significant advancements in image generation, with the diffusion model [21, 46] emerging as a leading method. This model has shown scalability with billion-scale web training data and has achieved remarkable quality in text-to-image generation tasks [5, 32, 35, 37, 39]. However, text-to-image models that rely solely on textual descriptions face limitations, particularly in scenarios requiring precise location control.

As the diversity and complexity of model training tasks increase, there is a growing demand for both accuracy and precision in generated data. Precision involves more accurate object positioning, while accuracy ensures that generated objects closely match intricate descriptions, even in highly complex scenarios. Recent approaches [25, 54] have addressed this by incorporating precise location control into the diffusion model, enabling open-vocabulary layout-to-image (L2I) generation. Among various layout types, bounding box based layouts offer intuitive and convenient control compared to masks or keypoints [39]. Additionally, bounding box layouts provide greater flexibility for diverse and detailed descriptions. In this work, we systematically investigate layout-to-image generation from bounding box-based layouts with rich context, where the descriptions for each instance to be generated can be complex, lengthy, and diverse, aiming to produce highly accurate objects with intricate and detailed descriptions.

Revisiting existing diffusion-based layout-to-image generation methods reveals that many rely on an extended self-attention mechanism [25, 54], which applies self-attention to the combined features of visual and textual tokens. This approach condenses textual descriptions for individual objects into single vectors and aligns them with image features through a dense connected layer.

However, a closer examination of how diffusion models achieve text-to-image generation [10, 32, 35, 37, 39] shows that text conditions are typically integrated via cross-attention layers rather than self-attention layers. Adopting cross-attention preserves text features as sequences of token embeddings instead of consolidating them into a single vector. Recent diffusion models have demonstrated improved generation results by utilizing larger [9, 10] or multiple [35] text encoders and more detailed image captions [5]. This underscores the significance of the cross-attention mechanism in enhancing generation quality through richer text representations and more comprehensive textual information.

Drawing an analogy between the generation of individual objects and the entire image, it is natural to consider applying similar cross-attention mechanisms to each object. Therefore, we propose introducing Regional Cross-Attention modules for layout-to-image generation, enabling each object to undergo a generation process akin to that of the entire image.

In addition to the proposed training scheme for L2I generation, we have identified a lack of reliable evaluation metrics for open-vocabulary L2I generation. While models [25, 54] can perform open-vocabulary L2I generation, evaluations are typically conducted on closed-set datasets such as COCO [7] or LVIS [18]. However, such closed-set evaluations may not accurately reflect the capabilities of open-vocabulary L2I models, as the text descriptions in these datasets are often limited to just a few words. It remains unclear whether these models can perform effectively when presented with complex and detailed object descriptions.

To address this gap, we propose two metrics that consider object-text alignment and layout fidelity that works for rich-context descriptions. Additionally, we conduct a user study to assess the reliability of these metrics and identify the circumstances under which these metrics may fail to reflect human preferences accurately.

Our contributions can be summarized as follows: 1) We revisit the training of L2I generative models and propose regional cross-attention module to enhance rich-context L2I generation, outperforming existing self-attention-based approaches. 2) To effectively evaluate the performance of open-set L2I models, we introduce two metrics that assess the models' capabilities with rich-context object descriptions and validate their reliability through a user study. 3) Our experimental results demonstrate that our proposed solution improves generation performance, especially with rich-context prompts, while reducing computational cost in each layout-conditioning layer thanks to the use of cross-attention.

https://doi.org/10.52202/079017-1983

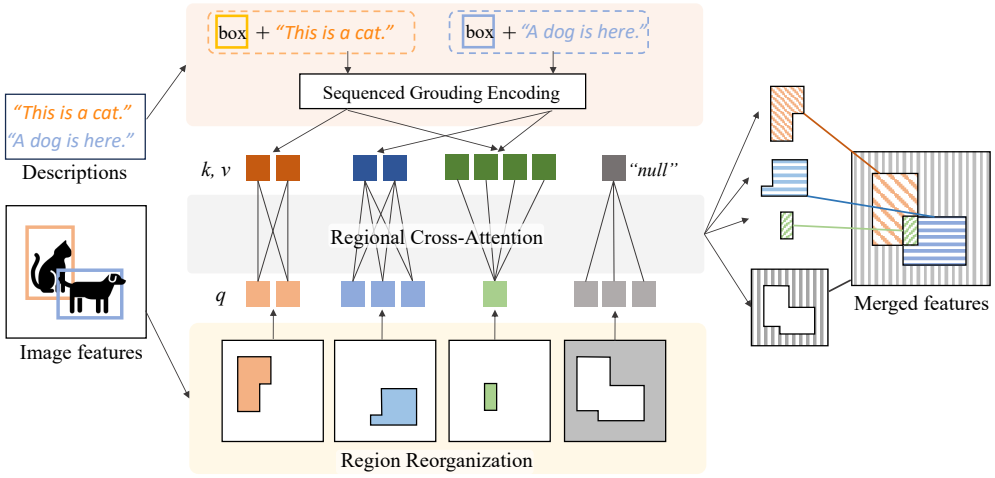


Figure 2: An example of regional cross-attention with two overlapping objects. Cross-attention is applied to each pair of regional visual and grounded textual tokens. The overlapping region cross-attends with the textual tokens containing both objects, while the non-object region attends to a learnable "$null$" token.

2 Related Works

Diffusion-based Generative Models The emergence of diffusion models [21, 46] has significantly advanced the field of image generation. Within just a few years, diffusion models have made remarkable progress across various domains, including super-resolution [44], colorization [41], novel view synthesis [55], 3D generation [14, 36, 51], image editing [6, 24, 29], image completion [42] and video editing [12, 45]. This progress can be attributed to several factors. Enhancements in network architectures [32, 34, 35, 37, 39, 43] have played a pivotal role. Additionally, improvements in training paradigms [15, 31, 47, 48] have contributed to this advancement. Moreover, the ability to incorporate various conditions during image generation has broadened the impact and applications of diffusion models. These conditions include elements such as segmentation [1, 2, 4, 59], using an image as a reference [16, 30, 40], and layout [11, 25, 54], the latter of which will be the main focus of our discussion in this work.

Layout-to-image generation: Early works [19, 23, 26, 33, 38, 49, 50, 60] often utilized GANs [17] or transformers [53] for L2I generation. For instance, GAN-based LAMA [26], LostGANs [49], and Context L2I [19] encode layouts as style features fed into adaptive normalization layers, while Taming [23] and TwFA [60] use transformers to predict latent visual codes from pretrained VQ-VAE [52]. Recent diffusion models [11, 25, 54, 56–58, 65] have shown promising results, extending L2I generation to be open-set. LayoutDiffuse [11] injects objects into the image features through learning per-class embeddings. LayoutDiffusion [65] fine-tunes pre-trained diffusion models by mapping object labels and layout coordinates into cross-attendable embeddings for attention layers. FreestyleL2I [58], BoxDiff [57] and R&B [56], which are training-free methods, leverage pre-trained diffusion models to inject objects into specified regions by imposing spatial constraints. GLIGEN [25] and InstDiff [54] explore open-set L2I generation using grounded bounding boxes, which encodes layout locations and object descriptions into features attended by self-attention layers.

3 Methodology

3.1 Challenges in Rich-Context Layout-to-Image Generation

The layout-to-image (L2I) generation task can be formally defined as follows: given a set of description tuples $S := \{s_i | s_i = (b_i, t_i)\}$, where b_i represents the bounding box coordinates of an object, and t_i denotes the corresponding text description, the objective is to generate an image that accurately aligns objects with their respective descriptions while maintaining fidelity to the specified layouts. In the closed-set setting, the number of text descriptions is limited to a fixed number N,

https://doi.org/10.52202/079017-1983

i.e., $N = |\{t_i\}|$, where N is the total number of classes. However, in the open-set and rich-context settings, the number of descriptions is unlimited, with descriptions in the rich-context setting being more diverse, complex, and lengthy.

Rich-context L2I encounters several challenges: 1) The rich-context descriptions for each object can be lengthy and complex, requiring the model to correctly understand the descriptions without overlooking details. Existing open-set layout-to-image solutions [25, 54] typically condense and map text embeddings into a single vector, which is then mapped to the image space for layout conditioning. However, this condensation process can result in significant information loss, particularly for lengthy descriptions. 2) Fitting various text descriptions into their designated layout boxes while maintaining global consistency is challenging. Unlike simpler text-to-image generation with a single description, L2I generation deals with multiple objects, requiring precise matching of each description to its specific layout area without causing global inconsistency. 3) L2I involves objects with intersecting bounding boxes, unlike segmentation-mask-to-image tasks where object areas do not overlap and can be efficiently handled by pixel-conditioned methods such as ControlNet [63] and Palette [42]. L2I models must determine the appropriate order and occlusion of overlapping objects autonomously, ensuring the proper representation and interaction of each object within the image.

3.2 Regional Cross-Attention

We propose using a regional cross-attention layer as a solution to rich-context layout-to-image generation, addressing the aforementioned challenges. The desired properties for an effective rich-context layout-conditioning module are as follows: 1) *Flexibility*: The model must accurately understand rich-context descriptions, regardless of their length or complexity, ensuring that no details are overlooked. 2) *Locality*: Each textual token should only attend to the visual tokens within its corresponding layout region, without influencing regions beyond the layout. 3) *Completeness*: All visual features, including those in the background, should be properly attended by certain description to maintain consistency in the output feature distribution. 4) *Collectiveness*: In cases where a visual token overlaps with multiple objects, it should consider all descriptions related to those intersecting objects.

Our approach differentiates itself from previous methods [25, 54] by employing cross-attention layers, rather than self-attention layers, to condition objects within the image. This design is inspired by the architecture of modern text-to-image diffusion models, which achieve fine-grained text control by incorporating pre-pooled textual features in the cross-attention layers. Analogously, one can apply cross-attention repeatedly between pre-pooled object description tokens and visual tokens within the corresponding regions for all objects. However, this straightforward method, though satisfies flexibility and locality, does not fully meet the criteria of completeness and collectiveness, as it may inadequately address non-object regions and overlapping objects. This limitation can result in inconsistent global appearances and challenges in managing overlapping objects effectively.

Region Reorganization. We propose region reorganization to satisfy locality, completeness and collectiveness, by creating a spatial partition of the image based on the layout. Each region is classified into one of three types: single-object region, overlapping region among objects, and background. This partitioning ensures that regions are mutually exclusive (*i.e.*, non-overlapping). Figure 2 illustrates a simple case with two overlapping objects. The overlapping area becomes a new, distinct region, while the non-overlapping parts of the original regions and remaining background are also treated as separate, new regions, thus ensuring completeness.

Formally, in the general case with multiple objects, the reorganized regions $R := \{r_i\}$ satisfy that the union of these regions will form a complete mask covering the entire visual feature space, while ensuring no intersection between any two reorganized regions:

$$\bigcup_{i=1}^{|R|} r_i = \mathbb{1}; \qquad r_i \cap r_j = \varnothing \quad \text{for } i \neq j \text{ and } i, j \in [1, |R|] \tag{1}$$

Our regional cross-attention operates within each reorganized region. We define a selection operation $f(\cdot, r_i)$ to identify the appropriate regions for cross-attention. For visual tokens $V := \{v_j\}$ it finds the tokens whose locations $\mathrm{loc}(v_j)$ lie within the i-th reorganized region. For description tuples S, it

filters the instances that overlap with the i-th reorganized region. This selection operation ensures that the text description is applied exclusively to the visual tokens within its corresponding region, thus maintaining locality. For regions with multiple objects, f also ensures that all overlapping descriptions are included to satisfy collectiveness.

$$f(V, r_i) := \{v_j | \text{loc}(v_j) \in r_i\} \tag{2}$$
$$f(S, r_i) := \{s_j | b_j \cap r_i \neq \varnothing\} \tag{3}$$

The final attention result A is the aggregation of all regional attention outputs. The selected descriptions $f(S, r_i)$ are encoded using Sequenced Grounding Encoding (SGE) in Figure 3 and serve as the key and value during cross-attention. For non-object regions where $f(S, r_i) = \varnothing$, a "$null$" embedding is learned as a substitute for the description.

$$a_i = \text{CrossAttn}(f(V, r_i), \text{SGE}[f(S, r_i)]); \qquad A = \bigcup_{i=1}^{|R|} a_i \tag{4}$$

Sequenced Grounding Encoding with Box Indicator. The selected object descriptions in each reorganized region are encoded into textual tokens using Sequenced Grounding Encoding in Figure 3. When multiple objects are present in a region, their descriptions are concatenated with a separator token. However, if two objects share the same description, their encoded textual embeddings will be identical. This identical embedding makes it impossible for the cross-attention module to distinguish between distinct objects. To address this issue, we incorporate bounding box coordinates as additional indicators. During encoding, we concatenate the bounding box coordinates with the textual tokens. The bounding box coordinates are initially encoded using sinusoidal positional encoding [53] and then repeated to match the length of the textual tokens before concatenation. For separator tokens and special tokens such as `[bos]` and `[eos]`, we use an all -1 vector for their box coordinates.

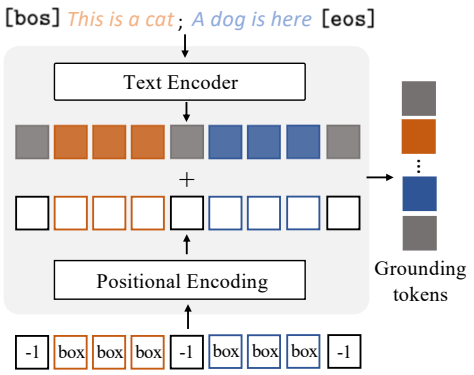


Figure 3: Sequenced Grounding Encoding with box coordinates as indicators.

4 Evaluation for Rich-Context L2I Generation

4.1 Rethinking L2I Evaluation

For evaluating layout-to-image generation, what are mainly considered are two aspects: 1) Object-label alignment, which checks whether the generated object matches the corresponding descriptions. 2) Layout fidelity, which examines how well the generated object aligns with the given bounding box.

In closed-set scenarios, it is common to use an off-the-shelf detector to evaluate L2I generation performance [11, 25, 54]. Object-label alignment is assessed by classifying image crops extracted from the generated image using a pre-trained classifier. Similarly, layout fidelity is measured by comparing the bounding boxes detected in the generated image with the provided layouts, using a pre-trained object detector.

However, in the open-set scenario, it is impossible to list all the classes. Moreover, even the state-of-the-art open-set object detectors [13, 61, 62] are set up to handle inputs at the word or phrase level, which falls short for the sentence-level descriptions required in rich-context L2I generation. we introduce two metrics to bridge the gap in evaluating open-vocabulary L2I models.

https://doi.org/10.52202/079017-1983

4.2 Metrics For Rich-Context L2I

We leverage the powerful visual-textual model CLIP for measuring object-label alignment, and the Segment Anything Model (SAM) for evaluating layout fidelity of the generated objects.

Crop CLIP Similarity: In rich-context L2I, object descriptions can be diverse and complex. The CLIP model, known for its robustness in image-text alignment, is thus suitable for this evaluation. To ensure accuracy and mitigate interference from surrounding objects, we compute the CLIP score after cropping the object as per the layout specifications.

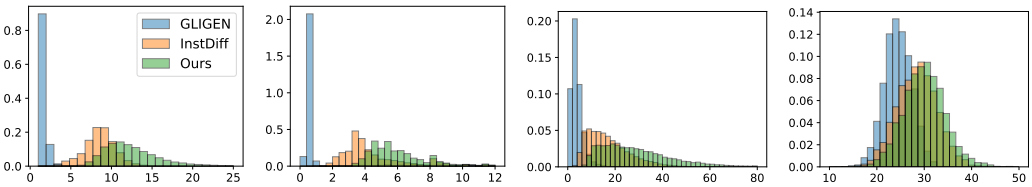
SAMIoU: An accurately generated object should closely align with its designated layout. Given the potential diversity in object shapes, we employ the SAM model, which can highlight an object's region in mask format within a given box region, to identify the actual region of the generated object. We then determine the generated object's circumscribed rectangle as its bounding box. The layout fidelity of the generated object with the ground-truth layout is quantified by the intersection-over-union (IoU) between the provided layout box and the generated object's circumscribed box.

5 Experiments

5.1 Model and Dataset

Model We leverage powerful pre-trained diffusion models as the foundation for our generative approach. Our best model is fine-tuned from Stable Diffusion XL (SDXL) [35]. We also provide the benchmarks using Stable Diffusion 1.5 (SD1.5) [39], which is a widely used backbone in existing methods [54]. The proposed regional cross-attention layer is inserted into the original diffusion model right after each self-attention layer. The weights of the output linear layer are initialized to zero, ensuring that the model equals to the foundational model at the very beginning. More implementation details is shown in Appendix B

Rich-Context Dataset To equip the model with the capability to be conditioned on complex and detailed layout descriptions, a rich-context dataset is required. While obtaining large-scale real-world datasets through human tagging is labor-intensive and expensive, synthetic training data can be more readily acquired by leveraging recent advancements in large visual-language models. Similar to GLIGEN [25] and InstDiff [54], we generate synthetic data to train our model.



(a) Average Caption Length (b) Gunning Fog Score (Complexity) (c) Unique Words/Sample (Diversity) (d) Object-label CLIP Alignment Score

Figure 4: Statistical comparisons between the synthetic object descriptions generated by GLIGEN [25], InstDiff [54], and our method. We measure the 1) average caption length, 2) the Gunning Fog Score, which estimates the text complexity from the education level required to understand the text, 3) the number of unique words per sample which indicates the text diversity, and 4) the object-label CLIP Alignment Score to measure object-label alignment. The results show that the pseudo-labels generated for our dataset are more complex, diverse, lengthier, and align better with objects, compared to those generated by GLIGEN and InstDiff.

We adopt a locating-and-labeling strategy during pseudo-label generation. At the first step, we use the Recognize Anything Model (RAM) [64] and GroundingDINO [27] to identify and locate salient objects in the image. Next, we use the visual-language model QWen [3] to produce the synthetic label for each object by asking it to generate detailed description of the object (see Appendix B for the prompt we used). We utilize CC3M [8] and COCO Stuff [7] as the image source. For COCO, we directly use the ground-truth bounding boxes rather than relying on RAM and GroundingDINO to generate synthetic labels. The final training dataset contains two million images, with 10,000 images from CC3M set aside and the 5,000-image validation set of COCO used for evaluation. We

https://doi.org/10.52202/079017-1983

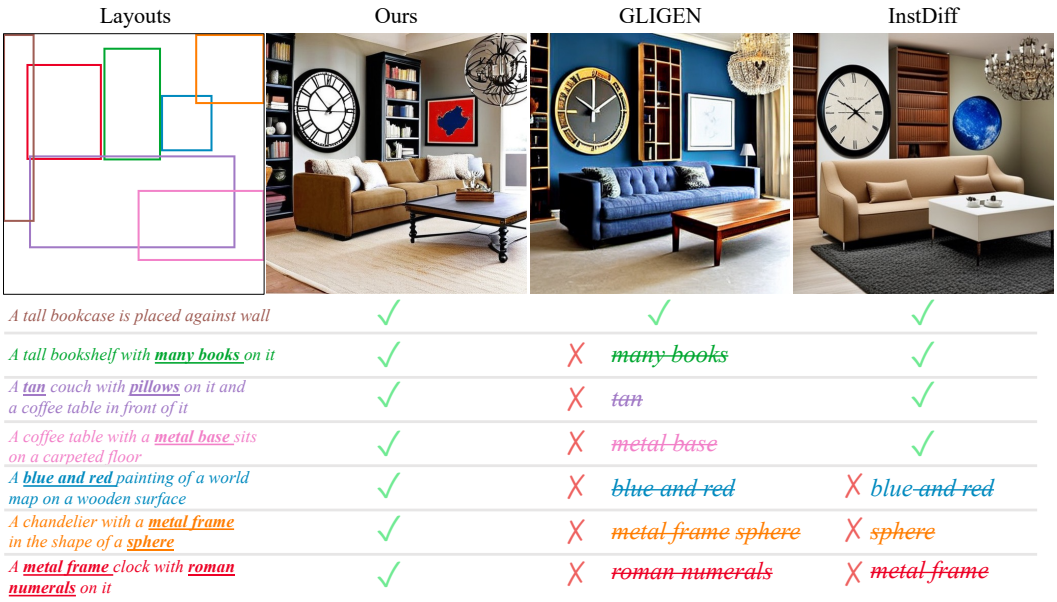


Figure 5: Qualitative comparison of rich-context L2I generation, showcasing our method alongside open-set L2I approaches GLIGEN [25] and InstDiff [54], based on detailed object descriptions. Our method consistently generates more accurate representations of objects, particularly in terms of specific attributes such as colors and shapes. Strikethrough text indicates missing content in the generated objects from the descriptions. More qualitative results available in Appendix H

denote the generated dataset Rich-Context CC3M (RC CC3M) and Rich-Context COCO (RC COCO). Compared to the synthetic training data used in GLIGEN [25] and InstDiff [54], our rich-context dataset provides more diverse, complex, lengthy, and accurate descriptions, as shown in Figure 4.

5.2 Reliability Analysis of Automatic L2I Evaluation

Our proposed evaluation metrics in Section 4.2 serve as a substitute for the lack of precise ground-truth in open-set scenarios. Whether the set is closed or open, the goal of evaluation is to ensure that the measurement results align with human perception. To validate the reliability of our automatic evaluation metrics, we conducted a user study on the RC CC3M dataset, randomly selecting 1000 samples. Each synthetic sample may contain multiple objects, but only one object was randomly selected for each question. Users were asked to answer two questions, each rated on a scale from 0 (bad) to 5 (good). For object-label alignment, users responded to the question: "Can the cropped object in the image be recognized as [label]?" with the label being the automatically generated object description from RC CC3M. For layout fidelity, users answered: "How well (tight) does the object align with the bounding box?" referring to the synthetic bounding box in RC CC3M.

In total, we collected 300 answers for each question. We used the Pearson correlation coefficient to analyze how well the automatic evaluation metrics align with human perception. The Pearson correlation coefficient measures the correlation between two distributions with a value ranging from -1 to 1, where 0 indicates no relation and 1 indicates a strong correlation. Empirically, we found that the automatic evaluation metrics sometimes failed to reflect human perception when the object size was very small or very large. We note that for small objects, the clarity of the object can be hampered, while large objects may overlap with many other objects, making the automatic measurements inaccurate. Therefore, we filtered out objects smaller than 5% or larger than 50% of the image, resulting in an improved Pearson correlation between automatic metrics and user scores from 0.33 to 0.59 for CropCLIP and from 0.15 to 0.52 for SAMIoU. We applied the same filtering rule in the remaining evaluations.

https://doi.org/10.52202/079017-1983

Table 1: Quantitative comparison of different L2I approaches under image resolution at 512x512. '↑' means that the higher the better, '↓' means that the lower the better.

	RC COCO			RC CC3M		
Method	CropCLIP↑	SAMIoU↑	FID↓	CropCLIP↑	SAMIoU↑	FID↓
Constrained						
BoxDiff [57]	22.61	58.75	29.99	-	-	-
R&B [56]	23.68	64.68	31.70	-	-	-
Open-Set						
GLIGEN [25]	25.20	78.66	27.62	25.27	83.64	<u>15.81</u>
InstDiff [54]	<u>27.46</u>	<u>80.78</u>	30.00	<u>28.46</u>	85.59	17.96
Ours						
SD1.5	27.36	<u>80.78</u>	**25.81**	28.45	<u>86.04</u>	16.49
SDXL	**28.15**	**80.84**	<u>27.41</u>	**29.42**	**86.56**	**11.02**

5.3 Rich-Context Layout-to-Image Generation

Evaluation Metrics. In addition to the two dedicated metrics discussed in Section 4.2, we also consider image quality by measuring the FID scores [20], which reflects how real or natural the generated images look compared to real images. While we did not observe significant changes in FID scores across different variations of our methods, we did notice change in image quality among different baseline methods.

Baseline Methods. We compare our approach with GLIGEN [25], a popular open-vocabulary L2I generative model, and InstDiff [54], a recent method that achieves state-of-the-art open-set L2I performance. Besides, two training-free L2I methods BoxDiff [57] and R&B [56] are also considered for comparison. Although these methods can accept open-set words, their inputs are limited to single words or simple phrases which are not truly rich-context descriptions. Therefore, we denote them as constrained L2I methods and only evaluate them on COCO using category names.

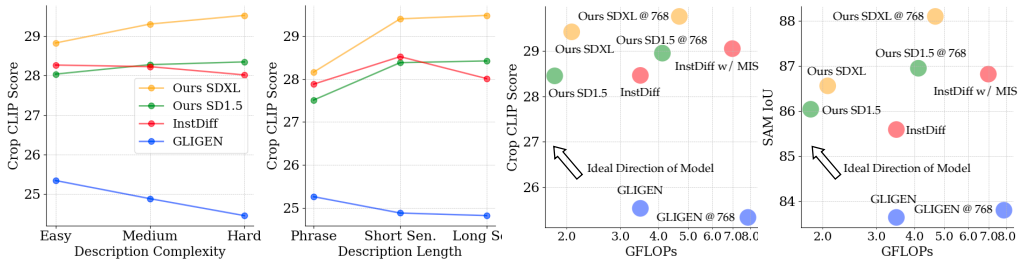
Table 1 benchmarks the performance of L2I methods at an image sampling resolution of 512. Our model with SD1.5 achieves similar performance to InstDiff while reducing the computation cost in the layout conditioning layer by half, as illustrated in Figure 6. Additionally, our model with SDXL achieves the best performance, even though the 512 resolution is sub-optimal for it. Further experiments in Section Section 5.5 demonstrate that higher sampling resolutions can further enhance performance. Figure 5 shows that, as the complexity and length of object descriptions increase, existing open-set L2I methods tend to overlook details especially when objects are specified with colors or shapes. In contrast, our method consistently generates objects that accurately represent the given descriptions.

5.4 Performance Across Various Complexity of Object Descriptions

By adopting pre-pooling textual features in the layout conditioning layer, our method maximizes the retention of textual information during generation. We observe that this design significantly enhances performance when dealing with complex and lengthy object descriptions. In Figure 6(a), we categorize object description complexity using the Gunning Fog score into three levels: easy (scores 0-4), medium (5-8), and hard ($>$8). Additionally, we classify descriptions by length into phrases ($\leq$8 words), short sentences ($\leq$15 words), and long sentences ($\geq$16 words). Our results indicate that for simple and short descriptions, the performance difference between our method and state-of-the-art open-set L2I methods is close. However, as the complexity and length of the descriptions increase, our method consistently outperforms existing approaches.

5.5 Ablation Study

We investigate the effectiveness of the proposed region reorganization and the use of bounding box indicators on object-label alignment and layout fidelity using RC CC3M dataset. In experiments without region reorganization, we use straightforward averaging features for overlapping objects. Empirically, we observe that without region reorganization, our model struggles to generate the



(a) (b)

Figure 6: (a) Object-text alignment scores across varying description complexities and lengths on RC CC3M. Our method shows significant advantages for complex and lengthy descriptions. (b) Object-text alignment and layout fidelity relative to computational cost in each layout-conditioning attention layer. Given that the number of textual tokens is much smaller than visual tokens, applying cross-attention can substantially reduce computational costs.

correct object when there is an overlap of objects with complex descriptions, leading to a significant drop in both object-label alignment and layout fidelity as shown in the Table 2.

Unlike self-attention-based solutions that use box indicators to implicitly indicate object locations, our method explicitly cross-attends visual tokens with corresponding textual tokens. This approach allows the model to recognize the correct location for object conditioning even without a box indicator. However, the reorganized mask in the regional cross-attention layer has a lower resolution than the original image, causing misalignment near the borders of generated objects. Adding a bounding box indicator not only helps the model distinguish objects with similar descriptions and but also improves layout fidelity, as validated by the improvement in SAMIoU.

Additionally, we notice that sampling at a higher image resolution (768x768) improves model performance, although it demands greater computational resources. It's important to note that generalization to higher resolution is not a universal capability of L2I models. Existing self-attention-based L2I methods like GLIGEN [25] experience performance declines when sampling at resolutions different from the training resolution. Another self-attention-based method, InstDiff [54], uses absolute coordinates for conditioning, requiring the sampling resolution to match the training resolution exactly. In Figure 6(b), we compare the performance-computation trade-off[2] of open-set L2I approaches. Since InstDiff does not support flexible resolution sampling, we utilize Multi-Instance Sampling (MIS) [54] instead. MIS was proposed to enhance InstDiff's performance by sampling each instance separately, albeit with increased inference times. We demonstrate the simplest case of MIS, which requires two inferences, but its computational cost scales linearly with the number of objects.

Table 2: Ablation study of the proposed methods on RC CC3M dataset with SDXL backbone. The results suggest that region reorganization plays an important role for rich-context L2I generation, while using box indicator and sample at higher-resolution can further enhance performance.

Region Reorg.	**Box Indicator**	**High Reso.**	CropCLIP↑	SAMIoU↑
✗	✗	✗	25.32	76.92
✓	✗	✗	28.93	85.02
✓	✓	✗	29.42	86.56
✓	✓	✓	29.79	88.10

6 Conclusion

In this study, we introduced a novel approach to enhance layout-to-image generation by proposing Regional Cross-Attention module. This module improve the representation of layout regions, particularly in complex scenarios where existing methods struggle. Our method reorganizes object-region correspondence by treating overlapping regions as distinct standalone regions, allowing for more accurate and context-aware generation. Additionally, we addressed the gap in evaluating open-vocabulary

[2]Computed using `https://github.com/MrYxJ/calculate-flops.pytorch` on an attention layer with 640 channels, which corresponds to a 32x32 resolution for image features, assuming input resolution of 512x512.

https://doi.org/10.52202/079017-1983

L2I models by proposing two novel metrics to assess their performance in open-set scenarios. Our comprehensive user study validated the consistency of these metrics with human preferences. Overall, our approach improves the quality of generated images, offering precise location control and rich, detailed object descriptions, thus advancing the capabilities of generative models in various potential applications.

Acknowledgement This work was funded in part by the Science and Technology Development Fund, Macau SAR (File no. 0049/2022/A1, 0050/2024/AGJ), and in part by the University of Macau (File no. MYRG2022-00072-FST, MYRG-GRG2024-00181-FST)

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023.

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[5] 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.

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[9] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$sigma$: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024.

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

[11] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023.

[12] Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. *arXiv preprint arXiv:2311.00213*, 2023.

[13] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. *arXiv preprint arXiv:2401.17270*, 2024.

[14] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023.

[15] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

[16] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[19] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15049–15058, 2021.

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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

[23] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. *arXiv preprint arXiv:2105.06458*, 2021.

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.

[25] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023.

[26] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13819–13828, 2021.

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[29] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[30] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[32] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[34] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[35] 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.

[36] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022.

 https://doi.org/10.52202/079017-1983

[37] 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.
[38] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.
[39] 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.
[40] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.
[41] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.
[42] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.
[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
[44] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.
[45] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.
[46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
[47] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.
[49] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10531–10540, 2019.
[50] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2647–2655, 2021.
[51] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023.
[52] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
[54] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. 2024.
[55] Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2022.
[56] Jiayu Xiao, Liang Li, Henglei Lv, Shuhui Wang, and Qingming Huang. R&b: Region and boundary aware zero-shot grounded text-to-image generation. *arXiv preprint arXiv:2310.08872*, 2023.
[57] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages

7452–7461, 2023.

[58] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14256–14266, 2023.

[59] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.

[60] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7764–7773, 2022.

[61] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022.

[62] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[63] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[64] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.

[65] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023.

 https://doi.org/10.52202/079017-1983

Supplementary Material for Rethinking The Training And Evaluation of Rich-Context Layout-to-Image Generation

A Limitations

Our work is built upon pre-trained diffusion models. Although our solution is backbone-agnostic, fine-tuning is still required when changing the model's backbone. Additionally, our training dataset is generated using visual-language model, which cannot guarantee that all synthetic labels are correct; these inaccuracies may negatively impact the model's performance. Furthermore, our training images are from publicly available datasets, which often contain low-quality images. As a result, the generated images may exhibit undesired artifacts, such as watermarks, during generation.

B Implementation Details

We train our model using the AdamW [28] optimizer with a learning rate of 5e-5. The training process involves an accumulated batch size of 256, with each GPU handling a batch size of 2 over 8 accumulated steps, for a total of 100,000 iterations on 16 NVIDIA V100 GPUs. This training process takes approximately 8,000 GPU hours. During training, we apply random cropping and horizontal flip for image augmentation, a bounding box will be dropped if its remaining size is smaller than 30% of its original size after cropping. We randomly drop 10% of layout conditions (all conditions in an image are dropped when a layout condition is dropped) and 10% of image captions to support classifier-free guidance [22]. During sampling, we use a classifier-free guidance scale of 4.5 for our SDXL-based model and 7.5 for our SD1.5-based model. The inference denoising step is set to 25 for our models and all baseline methods. During synthetic data generation, we obtain the description of the object using the following prompt for QWen model: "You are viewing an image. Please describe the content of the image in one sentence, focusing specifically on the spatial relationships between objects. Include detailed observations about all the objects and how they are positioned in relation to other objects in the image. Your response should be limited to this description, without any additional information".

C Throughput of Different Layout-to-Image Methods.

In addition to the FLOPs comparison presented in Section 5.5, we compare the throughput of using different L2I methods and present the result in the Figure 7

D Layout-to-Image Generation Diversity Comparison

Following LayoutDiffusion [65], we evaluate the generation diversity using LPIPS and Inception Score and present the diversity comparison of different L2I methods using 1,000 RC CC3M evaluation layouts in Table 3.

	LPIPS↑	**Inception Score↑**
GLIGEN	0.65 ± 0.10	15.74 ± 1.52
InstDiff	0.67 ± 0.10	15.26 ± 1.75
Ours (SD1.5)	$\mathbf{0.71} \pm 0.13$	$\mathbf{16.95} \pm 2.31$
Ours (SDXL)	$\underline{0.68} \pm 0.12$	$\underline{15.86} \pm 1.58$

Table 3: For LPIPS computation, each layout is inferred twice, and the score is calculated using AlexNet. A higher LPIPS score indicates a larger feature distance between two generated images with the same layouts, signifying greater sample-wise generation diversity. A higher Inception Score suggests a more varied appearance of generated images, indicating greater overall generation diversity.

https://doi.org/10.52202/079017-1983

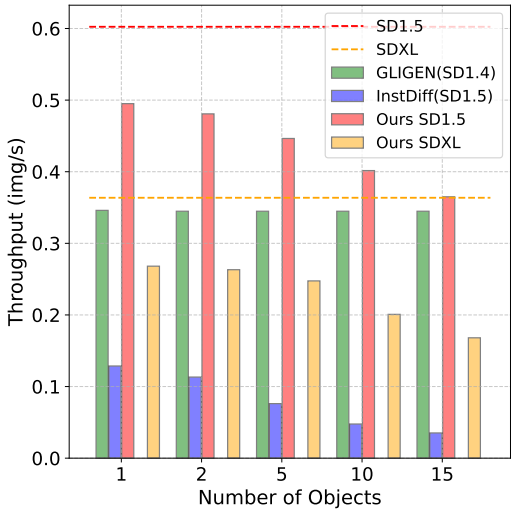


Figure 7: All methods are tested with float16 precision and 25 inference steps. The results are averaged over 20 runs. Notably, the overall throughput of our method is not significantly hampered. In a typical scenario with 5 objects, the throughput of our method exceeds 60% of the throughput of the original backbone model. Please note that while the official backbone of GLIGEN is SD1.4, its network structure and throughput are identical to those of SD1.5.

E Pseudo-code For Proposed Evaluation Metrics

Algorithms 1 and 2 shows the pseudo-code for computing the proposed Crop CLIP score and SAMIoU score on a single generated sample. The final performance is calculated by averaging these scores across all generated samples.

Algorithm 1 Compute Crop CLIP Score

Require: Generated image `I`, conditioning layout boxes `B` and labels `L` for each object, CLIP models clip_{img} and clip_{text}

```
crop_clip_scores ← []
for each (box, label) ∈ zip(B, L) do
    S ← crop(I, box)
    if S size < lower thres. or S size > upper thres. then
        continue
    end if
    clip_img_feat ← clip_img(S)
    clip_text_feat ← clip_text(label)
    crop_clip_sim ← cosine_similarity(clip_img_feat, clip_text_feat)
    crop_clip_scores.append(crop_clip_sim)
end for
sample_crop_clip_score ← mean(crop_clip_scores)
return sample_crop_clip_score
```

F Effectiveness of Rich-Context Dataset and Regional Cross-Attention

The ablation study in Table 4 indicates that to condition the L2I model with rich-context descriptions, both a rich-context dataset and a designated conditioning module for rich-context description are vital.

https://doi.org/10.52202/079017-1983

Algorithm 2 Compute SAMIoU Score

Require: Generated image `I`, conditioning layout boxes `B` for each object, Segment Anything Model SAM

```
sam_iou_scores ← []
for box ∈ B do
    if S size < lower thres. or S size > upper thres. then
        continue
    end if
    sam_mask ← SAM(I, b)
    box_of_generated_obj ← get_circumscribed_rectangle(sam_mask)
    sam_iou ← compute_IoU(box, box_of_generated_obj)
    sam_iou_scores.append(sam_iou)
end for
sample_sam_iou_score ← mean(sam_iou_scores)
return sample_sam_iou_score
```

Backbone	Dataset		Attention Module			CropCLIP	SAMIoU
	Word/Phrase	Rich-context	SelfAttn GLIGEN	SelfAttn InstDiff	CrossAttn Ours		
SDXL	✓				✓	25.40	86.76
SDXL		✓			✓	29.79	88.10
SD1.5		✓	✓			25.56	82.72
SD1.5		✓		✓		28.36	85.58
SD1.5		✓			✓	28.94	86.91

Table 4: The performance is evaluated on RC CC3M evaluation set and all methods are sampled under their best sampling resolution as discussed in Section 5.5. It can be noticed that even with the rich-context dataset, the performance of self-attention-based modules does not show significant improvement over their performance in the Table 1 in the paper.

G Effectiveness of Regional Cross-Attention with Visual Example

Figure 8 presents the visual comparison of using region reorganization compared to straightforward feature averaging when dealing with overlapping objects. The model with region reorganization can accurate generated objects that better align with the designated layouts, while the feature averaging solution can result in objects in incorrect location, generating undesired instances or making the overlapping instances inseparable.

H More Qualitative Results

During the evaluation, we filtered out very small and very large regions to avoid inconsistencies between automatic evaluations and human preferences. However, this does not imply that our method is incapable of generating high-quality small or large objects. Our results in Figure 9 verify that our model can accurately handle both very large and very small objects.

In addition, We provide more qualitative comparison with existing open-set L2I approaches in Figure 10.

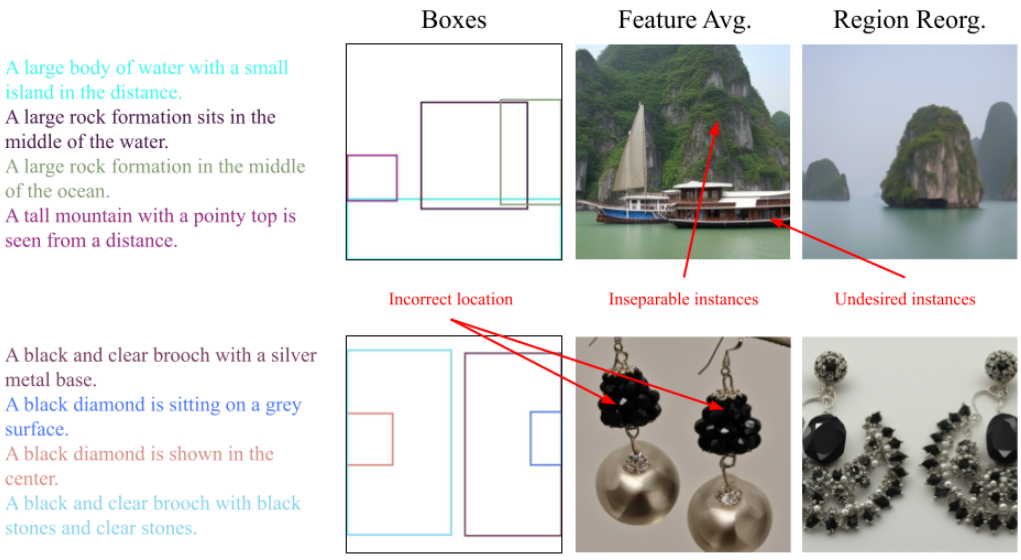


Figure 8: The model with region reorganization can accurate generated objects that better align with the designated layouts, while the feature averaging solution can result in objects in incorrect location, generating undesired instances or making the overlapping instances inseparable.

https://doi.org/10.52202/079017-1983

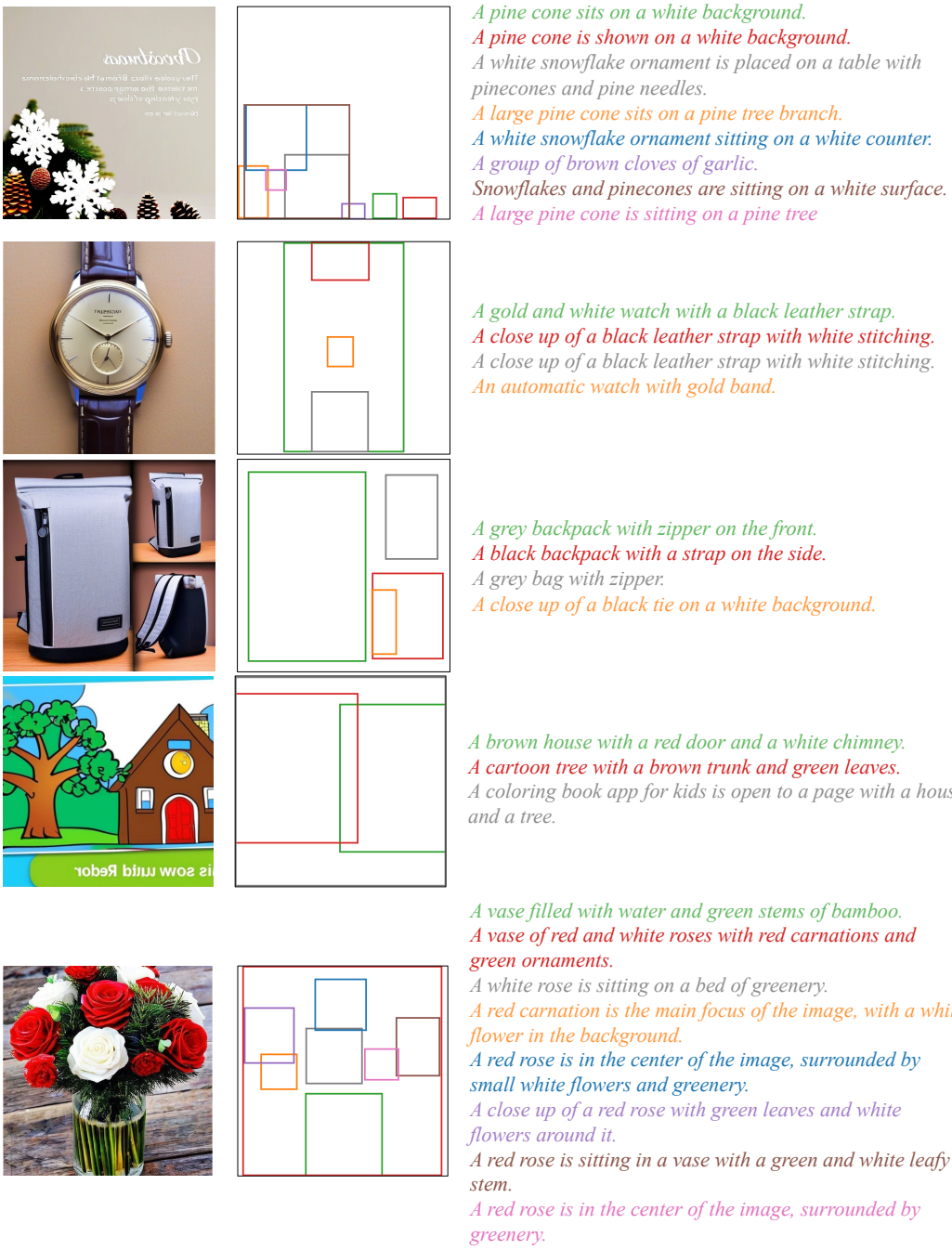


Figure 9: Additional qualitative results using random layouts from the synthetic RC CC3M dataset demonstrate our model's ability to accurately handle both very large and very small objects.

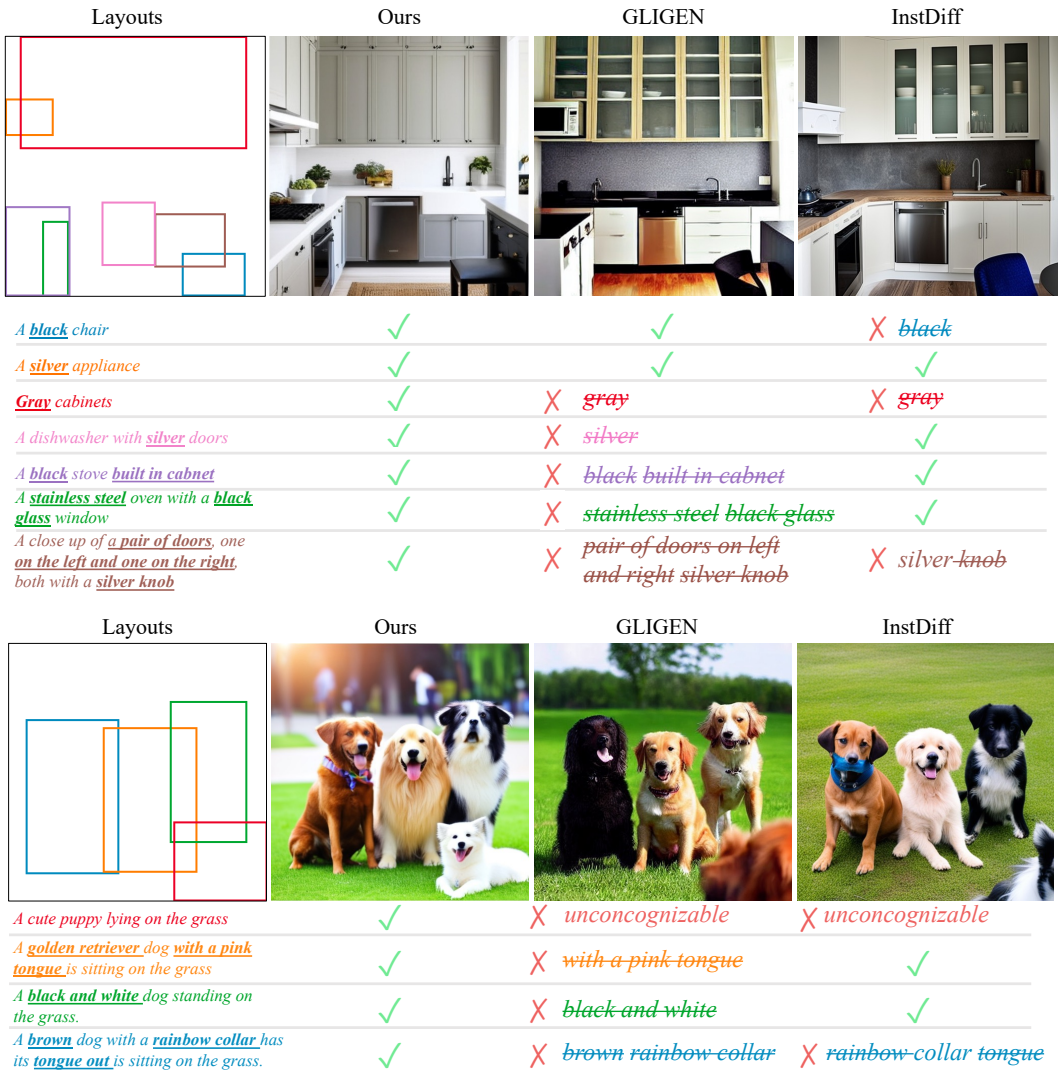


Figure 10: More qualitative comparison of rich-context L2I Generation, showcasing our method alongside open-set L2I approaches GLIGEN [25] and InstDiff [54], based on detailed object descriptions. Our method consistently generates more accurate representations of objects, particularly in terms of specific attributes such as colors and shapes. Strikethrough text indicates missing content in the generated objects from the descriptions.

https://doi.org/10.52202/079017-1983

1. **Claims**

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

   Answer: [Yes]

   Justification: We have conducted extensive experiments to validate our claims.

   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: Please see appendix

   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: [NA]

   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: We have provided the implementation details in the paper for result reproducibility.

   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]

https://doi.org/10.52202/079017-1983

Justification: Our code has been cleaned up and made public.

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: See implementation details

   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: [No]

   Justification: Our experiment is computationally intensive thus we only report the average number on large testing set.

   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: See implementation details

   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: We have checked the website.

   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: [No]

    Justification: On positive side, our work can provide better control for diffusion models, resulting in better application of generative models in research or industry. On negative side, it is possible our model will be misused for generating forgery content.

    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.

https://doi.org/10.52202/079017-1983

- 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: [NA]

    Justification: Our research do not have such risk.

    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 dataset and pretrained model used are all licensed for researched purpose.

    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: [NA]

    Justification: We are not releasing new assets.

    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: [Yes]

    Justification: We have provide the instruction for user study. Our use study is small scale with volunteers without compensation.

    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: Our study object is not human.

    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.

https://doi.org/10.52202/079017-1983