

Seg2Any: Open-set Segmentation-Mask-to-Image Generation with Precise Shape and Semantic Control

Danfeng Li^{1,2*} Hui Zhang^{1,2*} Sheng Wang³ Jiacheng Li³ Zuxuan Wu^{1,2†}

¹Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University

²Shanghai Collaborative Innovation Center of Intelligent Visual Computing

³HiThink Research

<https://seg2any.github.io>

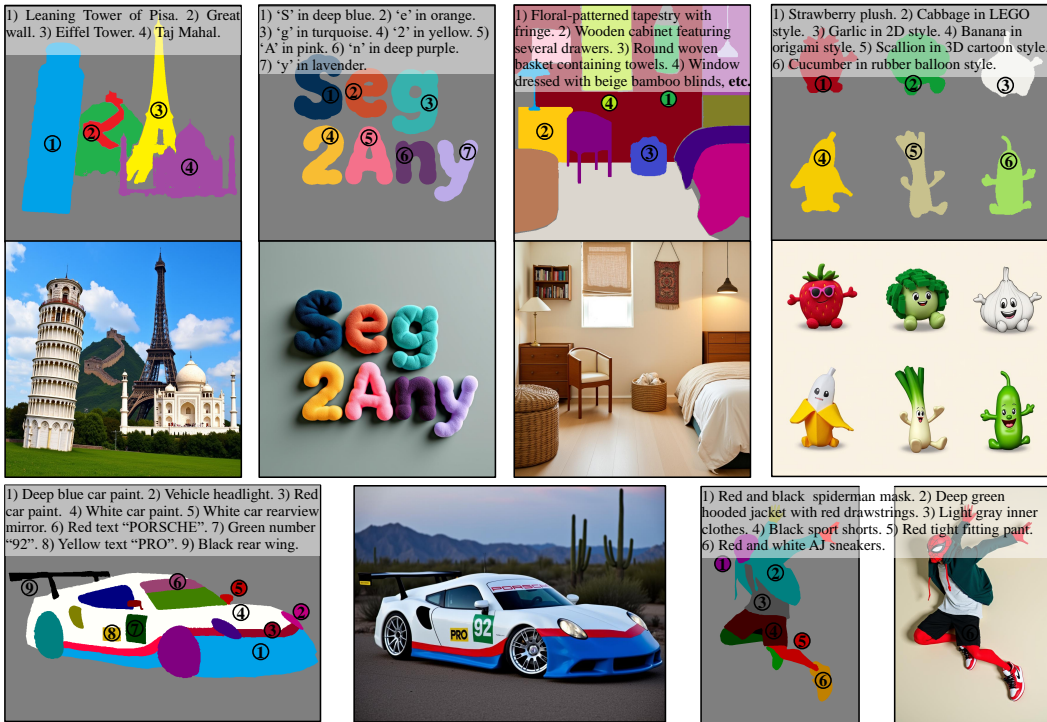


Figure 1: We propose Seg2Any, a novel segmentation-mask-to-image generation approach that achieves strong shape consistency and fine-grained attribute control (*e.g.* color, style, and text).

Abstract

Despite recent advances in diffusion models, top-tier text-to-image (T2I) models still struggle to achieve precise spatial layout control, *i.e.* accurately generating entities with specified attributes and locations. Segmentation-mask-to-image (S2I) generation has emerged as a promising solution by incorporating pixel-level spatial guidance and regional text prompts. However, existing S2I methods fail to simultaneously ensure semantic consistency and shape consistency. To address these challenges, we propose Seg2Any, a novel S2I framework built upon advanced multimodal diffusion transformers (*e.g.* FLUX). First, to achieve both semantic

*Equal contribution.

†Corresponding author.

and shape consistency, we decouple segmentation mask conditions into regional semantic and high-frequency shape components. The regional semantic condition is introduced by a Semantic Alignment Attention Mask, ensuring that generated entities adhere to their assigned text prompts. The high-frequency shape condition, representing entity boundaries, is encoded as an Entity Contour Map and then introduced as an additional modality via multi-modal attention to guide image spatial structure. Second, to prevent attribute leakage across entities in multi-entity scenarios, we introduce an Attribute Isolation Attention Mask mechanism, which constrains each entity’s image tokens to attend exclusively to themselves during image self-attention. To support open-set S2I generation, we construct SACap-1M, a large-scale dataset containing 1 million images with 5.9 million segmented entities and detailed regional captions, along with a SACap-Eval benchmark for comprehensive S2I evaluation. Extensive experiments demonstrate that Seg2Any achieves state-of-the-art performance on both open-set and closed-set S2I benchmarks, particularly in fine-grained spatial and attribute control of entities.

1 Introduction

Text-to-image (T2I) generation [34, 4, 9] has been widely adopted in various applications due to its powerful generative capabilities. However, T2I models struggle to achieve precise layout control (*i.e.* precisely generate entities in specified attributes and positions) solely through text prompts.

Layout-to-image generation has been proposed and designed to generate images based on specified layout conditions, including spatial locations and descriptions of entities. These layout conditions come in various forms, such as bounding boxes [22, 60, 44, 19, 54, 25], depth maps [62, 63, 61], segmentation masks [47, 24, 53], *etc.* This paper focuses on segmentation-mask-to-image (S2I) generation, where segmentation masks dictate the spatial locations of entities, and text descriptions specify their semantic content, thereby enabling the most fine-grained control over the images.

Existing S2I methods can be mainly divided into two categories: I) Methods that integrate segmentation masks as additional conditional inputs, such as ControlNet [56], ControlNet++ [20] and T2I-Adapter [26]. These methods often fail to align regional textual descriptions with their respective regions, resulting in semantic inconsistency in the generated images (see Figure 2 (a)). II) Methods based on the masked attention mechanism, which restricts each text embedding to attend solely to the respective image embeddings (*e.g.* FreestyleNet [47], PLACE [24] and EliGen [53]). Although these methods achieve semantic alignment, they fall short in precise shape preservation, as shown in Figure 2 (b) and (c). We attribute this shape inconsistency to the loss of spatial information when segmentation masks are compressed into the latent space. Notably, compared to the UNet [34] architecture with $8\times$ downsampling, recent advanced DiT [28] architectures employ a more aggressive $16\times$ downsampling, which further amplifies the loss of spatial information.

To address these challenges, we propose Seg2Any, a novel S2I framework built upon advanced multimodal diffusion transformers (*e.g.* FLUX [18]). Seg2Any mainly relies on two key innovations: I) **Semantic-Shape Decoupled Layout Conditions Injection**. We decouple segmentation mask conditions into two components: the shape condition (high-frequency shape information) and the regional semantic condition (low-frequency semantic information). *For the injection of regional semantic conditions, we employ a Semantic Alignment Attention Mask*. This mechanism tightly binds each entity’s image tokens to its corresponding text prompts, ensuring that the generated image is semantically consistent with the input descriptions at the entity level. *For the injection of shape condition, we propose Sparse Shape Feature Adaptation to integrate key spatial structure into the model efficiently*. Conventional approaches [56, 20] assign fixed colors to distinct categories in the semantic segmentation maps, which is only suitable for closed-set S2I generation and cannot be generalized to open-set scenarios. In contrast, we introduce an Entity Contour Map as our category-agnostic shape representation, which consists of entity contours extracted from segmentation masks. This representation is inherently sparse, preserving essential shape details in a compact and efficient manner. Following OminiControl [37], we integrate the condition tokens with text and noisy image tokens into a unified sequence, allowing them to interact directly through multi-modal attention [9]. As shown in Figure 2 (d), our approach achieves both semantic and shape consistency simultaneously.



Figure 2: Comparison in terms of shape and semantic consistency. Semantic inconsistency is annotated by blue boxes, while shape inconsistency is highlighted with red boxes, which reveal inconsistency in the number of vertical bars on the railings. In contrast, our approach achieves both shape and semantic consistency.

II) Attribute Isolation via Image Self-Attention Mask. Multi-instance generation often suffers from attribute leakage, where the visual attributes of one entity may affect others. To mitigate this problem, we introduce an Attribute Isolation Attention Mask strategy that prevents cross-entity information leakage by ensuring that image tokens associated with each entity are isolated from those of others.

Current dense caption datasets [33, 51, 31, 8, 50] are often limited by their closed-set vocabularies and coarse-grained descriptions, which constrain their effectiveness in training S2I models. Recent advances in open-source vision language models (VLMs), such as Qwen2-VL-72B [43], have significantly reduced the performance gap with close-source VLMs like GPT-4V [1], making it feasible to create large-scale and richly annotated datasets. Leveraging the capabilities of Qwen2-VL-72B, we construct Segment Anything with Captions 1 Million (SACap-1M), a large-scale dataset derived from the diverse and high-resolution SA-1B dataset [16]. SACap-1M contains 1 million image-text pairs and 5.9 million segmented entities, each comprised of a segmentation mask and a detailed regional caption, with captions averaging 58.6 words per image and 14.1 words per entity. We further present the SACap-Eval, a benchmark for assessing the quality of open-set S2I generation.

Through comprehensive evaluations on both open-set (SACap-Eval) and closed-set (COCO-Stuff, ADE20K) benchmarks, Seg2Any consistently outperforms prior SOTA models, particularly in fine-grained spatial and attribute control of entities.

To summarize, our contributions are as follows:

1. We propose Seg2Any, a framework that enables precise control over shape and semantics while preventing attribute leakage in open-set S2I generation.
2. We construct SACap-1M, a large-scale open-set dataset with 1M images and 5.9M regional annotations, along with SACap-Eval, an open-set benchmark for evaluating S2I generation.
3. Seg2Any achieves state-of-the-art performance on both open-set (SACap-Eval) and closed-set (COCO-Stuff, ADE20K) benchmarks.

2 Related Work

2.1 Text-to-Image Generation

Text-to-image (T2I) generation [34, 4, 9, 42, 38] has undergone significant advancement in recent years. Motivated by advances in large-scale transformer architectures, the Diffusion Transformer (DiT) [28] was introduced. Building upon this foundation, recent models like SD3 [9] and FLUX [18] further propose the Multimodal Diffusion Transformer (MM-DiT), which treats text as an independent modality and incorporates flow matching objectives, achieving state-of-the-art results. To integrate additional condition images (e.g., canny maps, depth maps, subject references) into MM-DiT, OminiControl [37] introduces a novel controllable framework. It concatenates condition image, text, and noisy image tokens, and then employs task-specific LoRA [13] modules to handle various conditions within a unified pipeline while maintaining minimal trainable parameters.

2.2 Layout-to-Image Generation

Layout-to-Image (L2I) generation is a task that synthesizes images guided by spatial layout conditions and entity-level textual descriptions. Existing L2I methods can be categorized according to the type of layout condition they employ. These include: **Bounding box-based approaches** [22, 60, 44, 19, 54, 55, 25, 11, 10, 36, 7, 58, 48] typically use rectangular spatial constraints to guide the image generation process, offering a coarse-grained layout control. **Depth map-based approaches** [62, 63, 61] utilize depth maps to achieve fine-grained spatial control similar to segmentation masks. For instance, 3DIS [62] divides multi-instance generation into two stages: it first generates a depth map via a text-to-depth model, then uses a pre-trained depth-to-image model to synthesize images with multi-instance attribute control. DreamRenderer [61], a training-free approach based on pre-trained controllable T2I models (FLUX.1-Depth [18] and FLUX.1-Canny [18]), identifies that middle layers in the FLUX model are responsible for instance-level rendering while shallow and deep layers capture global context. Consequently, it applies a hard image self-attention mask only to the middle layers to prevent attribute leakage. However, as a training-free approach, directly applying the attention mask severely impairs overall visual harmony.

Segmentation mask-based approaches [47, 24, 53, 27, 15, 2, 52] focus on pixel-level layout control. For example, FreestyleNet [47] uses binary attention weights in the cross-attention module that assigns a value 1 to allow text tokens to bind to corresponding image regions and 0 to prevent attention from unrelated areas. However, this approach requires downsampling the segmentation masks to align with the lower-resolution latent features, thereby sacrificing spatial detail. To alleviate the above issue, PLACE [24] introduces a layout control map that softens the attention weights. Instead of hard binary assignments in FreestyleNet, it calculates the area proportions of different entities within the receptive field of every image token, yielding soft attention weights. Yet, this approach struggles to mitigate spatial detail loss in MM-DiT architectures, where 16× downsampling (compared to 8× in U-Net) leads to more severe degradation of spatial information. The work most similar to ours is EliGen [53], which is also built on FLUX. Unlike our approach, it is trained on datasets with bbox annotations and only supports loose position control through scribble-style masks, whereas our approach enables both strict and loose mask position control. Additionally, it relies solely on a masked attention mechanism without injecting explicit spatial information.

3 Methodology

3.1 Problem Definition

We define the instruction y for a segmentation-mask-to-image model as a composition of a global text prompt and N entity-level text prompts with corresponding binary masks:

$$y = [p_0, (p_1, m_1), \dots, (p_i, m_i), \dots, (p_N, m_N)], i \in [1, N], \quad (1)$$

where p_0 denotes the global textual description, while each p_i represents the entity-specific textual prompt with its corresponding binary segmentation mask m_i for the i -th entity.

3.2 Semantic-Shape Decoupled Layout Conditions Injection

As shown in Figure 3, we decouple segmentation mask-based layout conditions into complementary semantic and shape components. For semantic information, we employ a **Semantic Alignment Attention Mask** (Section 3.2.1) mechanism that binds text prompts to their corresponding image regions. For shape information, we adopt the **Sparse Shape Feature Adaptation** (Section 3.2.2) to efficiently encode and integrate spatial layout conditions into the model.

3.2.1 Semantic Alignment Attention Mask

To ensure semantic alignment between textual and visual modalities, we introduce a **Semantic Alignment Attention Mask**, denoted as $\mathbf{M}_{\text{sem-align}}$, which governs the interactions between textual and visual tokens. We denote the text token indices of the global caption as \mathcal{T}_0 , and those of the i -th entity’s regional caption as \mathcal{T}_i ($i = 1, \dots, N$). Likewise, \mathcal{I}_0 and \mathcal{I}_i correspond to the image token indices of the background and the i -th entity, respectively. The $\mathbf{M}_{\text{sem-align}}$ is then defined as follows:

$$\mathbf{M}_{\text{sem-align}}[q, k] = \begin{cases} 1, & \text{if } q \in \mathcal{T}_i, k \in \mathcal{T}_i \quad (\text{text-text}) \\ 1, & \text{if } q \in \mathcal{T}_i \cup \mathcal{T}_0, k \in \mathcal{I}_i \quad (\text{text-image}) \\ 1, & \text{if } q \in \mathcal{I}_i, k \in \mathcal{T}_i \cup \mathcal{T}_0 \quad (\text{image-text}) \\ 1, & \text{if } q, k \in \bigcup_{i=0}^N \mathcal{I}_i \quad (\text{image-image}) \\ 0, & \text{otherwise} \end{cases}, i \in [0, N], \quad (2)$$

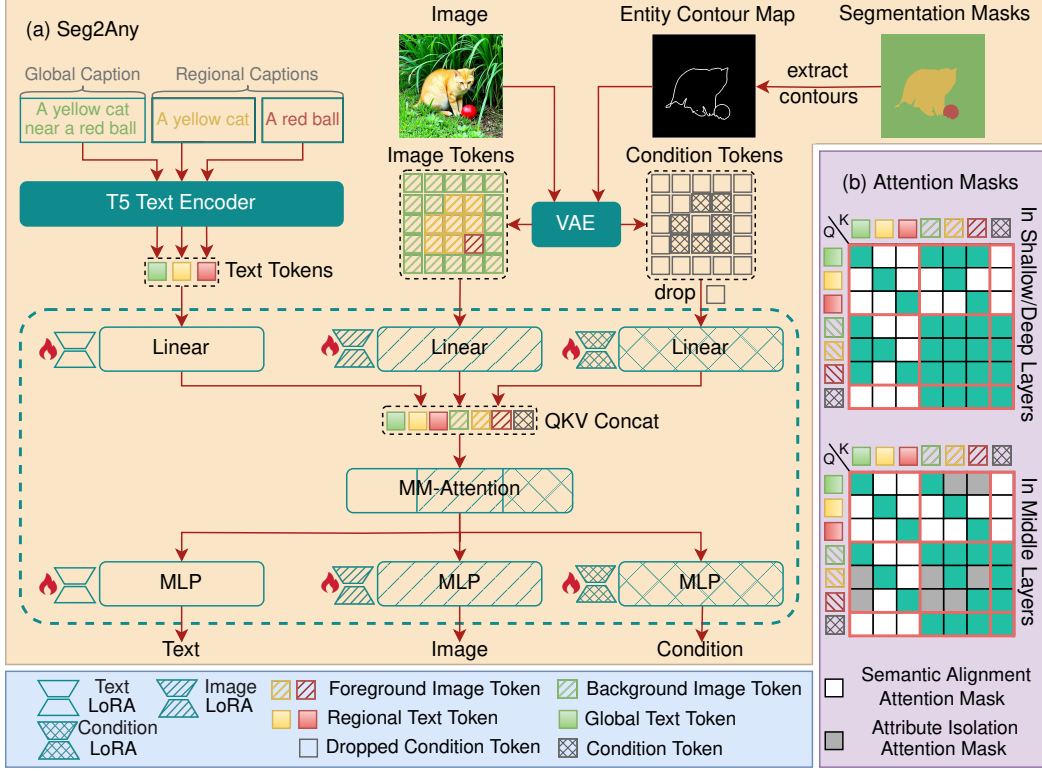


Figure 3: (a) An overview of the Seg2Any framework. Segmentation masks are transformed into Entity Contour Map, then encoded as condition tokens via frozen VAE. Negligible tokens are filtered out for efficiency. The resulting text, image, and condition tokens are concatenated into a unified sequence for MM-Attention. Our framework applies LoRA to all branches, achieving S2I generation with minimal extra parameters. (b) Attention Masks in MM-Attention, including Semantic Alignment Attention Mask (Section 3.2.1) and Attribute Isolation Attention Mask (Section 3.3).

where q and k are the query and key indices, respectively. As illustrated in Figure 3 (b), this attention mechanism guarantees that each generated entity adheres closely to its text prompt. Meanwhile, all image tokens attend to each other to ensure globally coherent visual synthesis. The global caption tokens \mathcal{T}_0 are allowed to attend to all image tokens, providing contextual global guidance.

3.2.2 Sparse Shape Feature Adaptation

Condition Image Representation. In open-set S2I generation, the common practice of using fixed, class-specific colors to represent semantic segmentation maps is inherently limited. To address this limitation, we use an **Entity Contour Map** to effectively encode shape information.

Starting with a set of binary masks $\{m_i\}_{i=1}^N$ for N distinct entities where each $m_i \in \{0, 1\}^{H \times W}$, we extract the contour of each mask as $\text{Contour}(m_i) \in \{0, 1\}^{H \times W}$. These contours are then merged into a single binary map:

$$C_{\text{gray}}(x, y) = \max_{1 \leq i \leq N} \text{Contour}(m_i)(x, y). \quad (3)$$

The resulting grayscale map $C_{\text{gray}} \in \mathbb{R}^{H \times W}$ is then further converted to an RGB image by duplicating the gray channel across all three channels, resulting in our Entity Contour Map $C \in \mathbb{R}^{H \times W \times 3}$.

This sparse shape representation offers advantages in our framework, as semantic information is already integrated through the Semantic Alignment Attention Mask mechanism (Section 3.2.1). This approach eliminates the need for dense, per-pixel semantic labels to indicate region occupancy. Due to its efficiency and sparsity, we adopt this representation as the shape encoding method in our work.

Minimal Condition Image Control. Inspired by OminiControl [37], we treat the Entity Contour Map—a type of image-based condition—as an independent modality and leverage LoRA [13] to

minimize training overhead. As shown in Figure 3 (a), the Entity Contour Map is encoded into condition image tokens by a frozen VAE encoder, concatenated with text and noisy image tokens to form the joint input sequence, which directly participates in multi-modal attention. Furthermore, the condition tokens share the same 2D position indices with the noisy image tokens under the RoPE encoding, which helps preserve spatial alignment.

Notably, unlike OminiControl, which applies LoRA only to the condition branch. We apply LoRA to all branches (as shown in Figure 3 (a)). The condition branch is trained with LoRA to seamlessly incorporate the Entity Contour Map. Meanwhile, the image and text branches are also trained using LoRA to ensure precise alignment between the generated entities and the regional text prompts. This approach modifies the linear layers in each DiT block across all three branches as follows:

$$\begin{aligned} W_{\text{cond}}^{\text{new}} &= W_{\text{img}} + B_{\text{cond}} A_{\text{cond}}, \\ W_{\text{img}}^{\text{new}} &= W_{\text{img}} + B_{\text{img}} A_{\text{img}}, \\ W_{\text{text}}^{\text{new}} &= W_{\text{text}} + B_{\text{text}} A_{\text{text}}, \end{aligned} \quad (4)$$

where $A_{\text{cond}}, A_{\text{img}}, A_{\text{text}} \in \mathbb{R}^{r \times d}$ and $B_{\text{cond}}, B_{\text{img}}, B_{\text{text}} \in \mathbb{R}^{d \times r}$ are the low-rank adaptation matrices for each respective branch ($r \ll d$). Here, $W_{\text{img}}, W_{\text{text}} \in \mathbb{R}^{d \times d}$ denote the original weight matrices of the linear layers.

Shape Guidance Strength Modulation. To adjust the influence of condition image tokens, the attention mechanism is adapted by incorporating a bias term [37], defined as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d}} + \text{Bias}(\gamma) \right) V, \quad (5)$$

where Q, K, V are computed from the concatenation of text tokens, noisy image tokens, and condition tokens. The bias matrix $\text{Bias}(\gamma) \in \mathbb{R}^{(L_{\text{text}}+2L_{\text{img}}) \times (L_{\text{text}}+2L_{\text{img}})}$, with L_{text} and L_{img} denoting the number of text and image-related tokens respectively, is given by:

$$\text{Bias}(\gamma) = \begin{bmatrix} 0_{L_{\text{text}} \times L_{\text{text}}} & 0_{L_{\text{text}} \times L_{\text{img}}} & 0_{L_{\text{text}} \times L_{\text{img}}} \\ 0_{L_{\text{img}} \times L_{\text{text}}} & 0_{L_{\text{img}} \times L_{\text{img}}} & \log(\gamma) \cdot \mathbf{1}_{L_{\text{img}} \times L_{\text{img}}} \\ 0_{L_{\text{img}} \times L_{\text{text}}} & \log(\gamma) \cdot \mathbf{1}_{L_{\text{img}} \times L_{\text{img}}} & 0_{L_{\text{img}} \times L_{\text{img}}} \end{bmatrix}. \quad (6)$$

The factor $\gamma \in (0, 1]$ serves to modulate the strength of shape conditioning. As γ approaches 1, $\log(\gamma)$ tends to 0, and the condition tokens retain full influence, enforcing strict adherence to the shape guidance. In contrast, as γ approaches 0, $\log(\gamma)$ tends to $-\infty$, which suppresses the contribution of condition tokens, thus making scribble-style segment masks feasible for flexible control.

Condition Image Token Filtering. Given the sparse nature of the Entity Contour Map, where numerous areas exhibit low or zero values, we discard tokens that are entirely composed of zero values, as they provide no shape information. This results in a significant reduction in tokens without compromising shape details. The token filtering process is illustrated in Figure 3 (a).

3.3 Attribute Isolation Attention Mask

A critical challenge in multi-instance generation is attribute leakage, where visual attributes from one entity transfer to others (as illustrated in Figure 4). To address this, we introduce the **Attribute Isolation Attention Mask**, defined as $\mathbf{M}_{\text{attr-isolate}}$:

$$\mathbf{M}_{\text{attr-isolate}}[q, k] = \begin{cases} 1, & \text{if } q \in \mathcal{T}_i, k \in \mathcal{T}_i \quad (\text{text-text}) \\ 1, & \text{if } q \in \mathcal{T}_i, k \in \mathcal{I}_i \quad (\text{text-image}) \\ 1, & \text{if } q \in \mathcal{I}_i, k \in \mathcal{T}_i \quad (\text{image-text}) \\ 1, & \text{if } q \in \mathcal{I}_i \cup \mathcal{I}_0, k \in \mathcal{I}_i \quad (\text{image-image}) \\ 0, & \text{otherwise} \end{cases}, i \in [0, N], \quad (7)$$

Unlike the Semantic Alignment Mask (Eq. 2), the Attribute Isolation Mask operates with stricter constraints, as illustrated in Figure 3 (b). First, it prevents cross-entity visual information interaction by restricting each entity’s image tokens to attend only to themselves. Second, it restricts the global caption tokens (\mathcal{T}_0) from attending to any foreground image tokens. This complete separation of entities effectively prevents attribute leakage. Notably, background image tokens are still permitted to attend to all image tokens, ensuring environmental coherence.

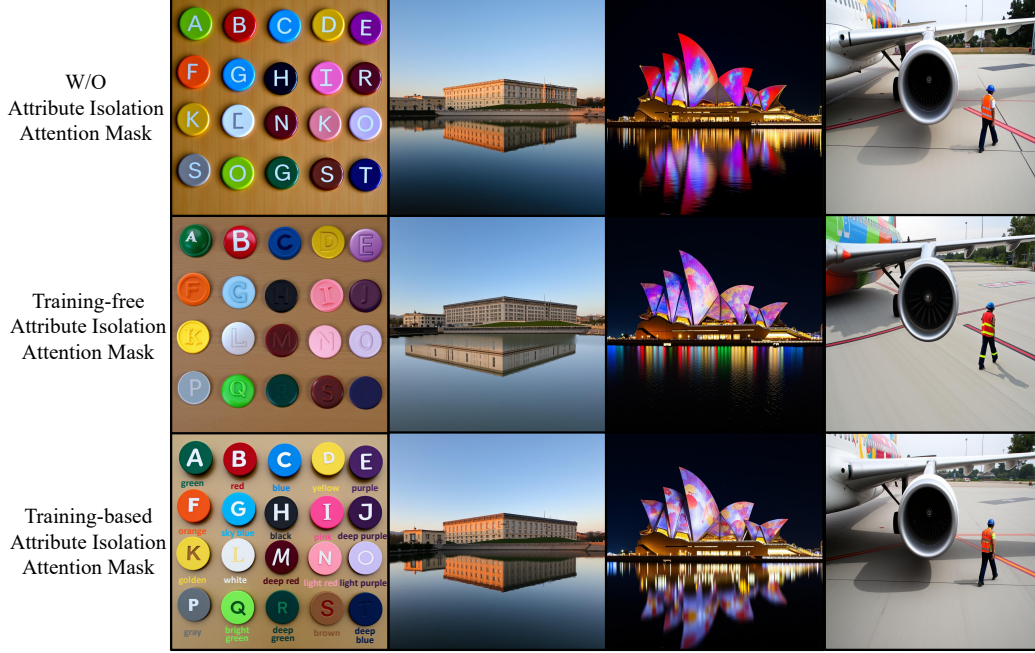


Figure 4: Visualization results of different attribute isolation strategies. In Column 1, 20 colored circular badges labeled A to T are required to be generated in raster order. The results show that our Attribute Isolation Attention Mask effectively prevents attribute leakage between entities. Columns 2-4 demonstrate that direct application of the mask without training leads to visual inconsistencies, manifesting as unnatural shadows and reflections. In contrast, the training-based approach on our proposed large-scale dataset achieves both strong attribute control and high visual coherence.

Building on insights from DreamRender [61] that the middle layers (20-38 layers) of the 57-layer FLUX architecture are dedicated to processing the visual features of individual instances, we apply the Attribute Isolation Attention Mask to these middle layers. In practice, we observe that training-free rigid attention constraints often introduce visual artifacts, as demonstrated in Figure 4. Instead, through training on our proposed large-scale datasets, deeper layers learn to refine holistic image quality, achieving an optimal balance between visual harmony and attribute control.

3.4 SACap-1M Dataset

To address the lack of large-scale and fine-grained datasets for S2I generation, we construct SACap-1M, containing 1 million image-text pairs and 5.9 million segmented entities with detailed descriptions. We propose an automated pipeline for data annotation and filtering: I) **Image Filtering**. We select images from the high-resolution and wide-ranging SA-1B [16] dataset. Initially, we remove images that are excessively large or small in size, as well as those with extreme aspect ratios. Subsequently, we apply the LAION-Aesthetics predictor [35] to filter out low-quality images with an aesthetic score below 5. II) **Entity Extraction**. The SA-1B dataset provides accurate, category-agnostic masks for each image. However, each image contains on average over 100 masks, many of which are nested. To filter masks, we retain only top-level masks by removing those that are contained within top-level masks. Additionally, we discard masks whose area is smaller than 1% of the total image area. Finally, we exclude images whose number of remaining masks falls outside the range of 1 to 20. III) **Regional and Global Caption Annotation**. We employ the open-source Qwen2-VL-72B [43] model to generate captions for each entity and the global image, yielding an average of 58.6 words per image and 14.1 words per entity. See supplemental materials for more details.

4 Experiments

4.1 Experiment Setup

Baselines. Our method is compared with prior state-of-the-art S2I approaches, including FreestyleNet [47] and PLACE [24]. Controllable T2I models, such as ControlNet [56], ControlNet++ [20], and UniControl [29], are also considered. Additionally, recent FLUX-based models (DreamRender [61], 3DIS [63], and EliGen [53]) are included in the comparison. Notably, DreamRender and 3DIS are training-free methods that require depth maps as input instead of segmentation masks. For evaluation of these models, we use Depth Anything V2 [49] to predict depth maps from real images as guidance.

Training datasets and Evaluation Benchmarks. Experiments are conducted on both open-set and closed-set segmentation datasets. *For the open-set scenario*, we utilize SACap-1M, which consists of 1 million images accompanied by 5.9 million regional captions. Evaluation for this setting is performed on SACap-Eval, a benchmark curated from a subset of SACap-1M, comprising 4,000 prompts with detailed entity descriptions and corresponding segmentation masks, with an average of 5.7 entities per image. *For closed-set scenario*, we select two widely used datasets: ADE20K [59] and COCO-Stuff [3]. Following prior works [47, 24], for ADE20K and COCO-Stuff, regional captions are assigned as the semantic class names of each segment, and no global caption is provided.

Implementation Details. Our experiments are conducted based on FLUX.1-dev. The LoRA modules are applied to all linear layers of each block in DiT, with the LoRA rank set to 64, resulting in approximately 594M additional parameters. Across all datasets, our model is trained for 20,000 steps with a batch size of 16, using the AdamW optimizer and a fixed learning rate of 0.0001. The training resolution is set to 1024×1024 for the SACap-1M dataset and 512×512 for ADE20K and COCO-Stuff. All experiments are conducted on 4 NVIDIA H100 GPUs.

Evaluation Metrics. *For closed-set S2I generation*, we report both FID and MIoU. FID reflects the visual fidelity of the generated images, while MIoU measures semantic and layout consistency. For MIoU calculation, we use Mask2Former [6] for ADE20K and DeepLabV3 [5] for COCO-Stuff to predict semantic segmentation, as done in ControlNet++ [20]. *For open-set S2I generation*, following CreatiLayout [54], we evaluate S2I quality on SACap-Eval from three perspectives:

- **Class-agnostic MIoU.** We use the ground-truth segmentation masks as mask prompts, and then employ SAM2 [32] to predict class-agnostic segmentation masks for the generated images. The predicted masks are compared with the ground truth to compute the class-agnostic MIoU, which measures shape consistency.
- **Region-wise quality.** Image regions are cropped based on ground-truth segmentation masks and evaluated with Qwen2-VL-72B [43] in a Visual Question Answering (VQA) manner to measure both spatial and attribute accuracy. Specifically, spatial accuracy is evaluated by checking whether each entity appears within the correct region, while attribute accuracy considers whether its color, text, and shape match the provided descriptions.
- **Global-wise quality.** We assess overall visual quality and global caption fidelity using multiple metrics, including the scoring models IR score [46] and Pick score [17], as well as commonly used metrics such as CLIP score and FID.

4.2 Quantitative Results

Fine-Grained open-set S2I. Table 1 presents the quantitative results on the SACap-Eval benchmark. We report multiple metrics covering class-agnostic shape consistency, region-wise, and global-wise qualities. Compared to previous methods, Seg2Any achieves significant improvements across most criteria. Notably, our approach reaches a class-agnostic MIoU of 94.90, nearly matching the real image upper bound (96.03 class-agnostic MIoU), which demonstrates excellent shape consistency. In terms of region-wise assessment, our method excels in region-wise spatial localization and precise attribute control, effectively mitigating attribute leakage. The qualitative results presented in Figure 5 further support this observation. For global quality, our method achieves strong overall performance thanks to the FLUX model and high-quality training data.

Table 1: Quantitative comparison on the SACap-Eval benchmark. **Bold** and underline represent the best and second best methods, respectively.

Method	Class-agnostic MIoU \uparrow	Region-wise Quality				Global-wise Quality			
		Spatial \uparrow	Color \uparrow	Shape \uparrow	Texture \uparrow	IR [46] \uparrow	Pick [17] \uparrow	CLIP \uparrow	FID \downarrow
Real Images	96.03	97.04	93.87	91.66	92.50	-	-	-	-
FreestyleNet [47]	74.59	42.34	40.08	25.07	40.07	-1.96	18.13	19.66	46.20
PLACE [24]	84.30	79.05	49.40	52.00	57.96	-1.11	19.71	24.69	17.81
EliGen [53]	51.38	83.62	76.94	78.91	77.21	0.49	22.60	27.57	19.69
3DIS [63]	82.09	88.25	85.80	85.87	<u>89.20</u>	0.53	<u>21.81</u>	<u>28.03</u>	<u>15.36</u>
DreamRender [61]	82.32	<u>88.69</u>	<u>87.73</u>	<u>86.76</u>	89.19	0.43	21.58	28.21	13.71
Seg2Any (ours)	94.90	93.89	91.52	88.15	90.12	0.44	21.66	27.87	15.53

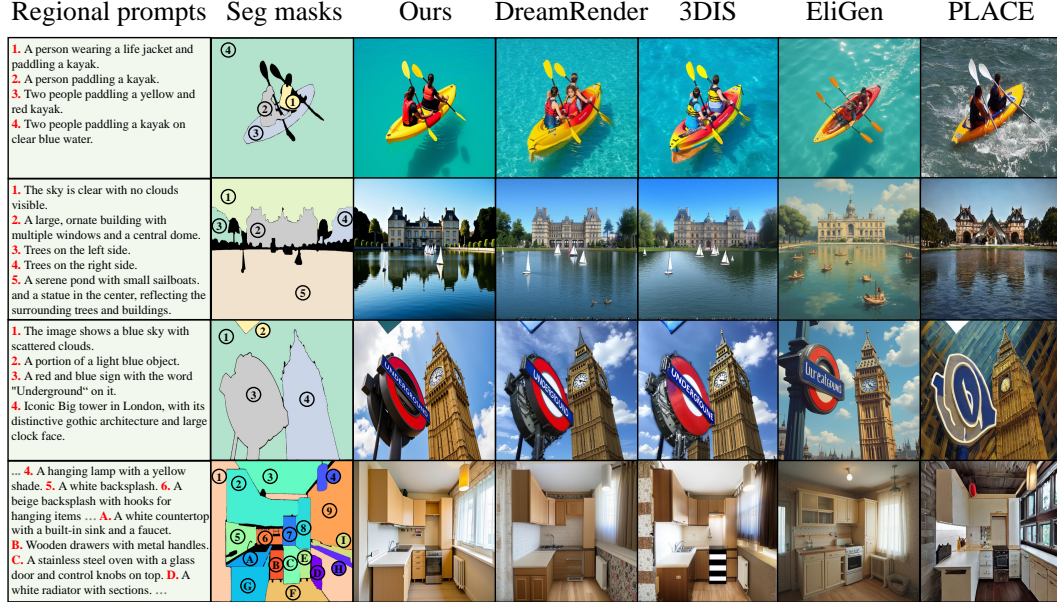


Figure 5: Qualitative comparisons on SACap-Eval. Seg2Any accurately generates entities exhibiting complex attributes such as color and texture, surpassing previous approaches.

Coarse-Grained Closed-set S2I. Table 2 shows quantitative results for closed-set S2I generation on the COCO-Stuff and ADE20K datasets. On COCO-Stuff, Seg2Any achieves the highest MIoU. On ADE20K, Seg2Any reaches 54.46 MIoU, approaching the upper bound of real images (54.41 MIoU). Although Seg2Any does not surpass PLACE (60.20 MIoU) on ADE20K, it outperforms all other baselines, demonstrating strong overall performance across datasets. Seg2Any exhibits higher FID scores compared to PLACE and FreestyleNet, which we attribute to the domain gap between the base model FLUX and the real images in ADE20K and COCO-Stuff.

4.3 Ablation Study

We conduct an ablation study to evaluate the individual contributions of our key components on the SACap-Eval benchmark (see Table 4). Introducing explicit shape condition via Sparse Shape Feature Adaptation (SSFA) significantly improves the class-agnostic MIoU metric over using only the Semantic Alignment Attention Mask (SAA), confirming the effectiveness of preserving shape details. Further, adding the Attribute Isolation Attention Mask (AIA) notably improves region-wise quality, indicating better prevention of attribute leakage. The training-free version degrades global quality; however, after training on our large-scale dataset, the overall image quality improves significantly. Table 4 also shows that incorporating Condition Image Token Filtering (CITF) leaves performance almost unchanged; its impact on computational cost is further analyzed in the supplemental materials.

On COCO-Stuff and ADE20K, we observe a similar trend: SSFA substantially improves MIoU, and CITF shows a marginal impact on performance. (see Table 3).

Table 2: Quantitative evaluation of S2I models on COCO-Stuff and ADE20K.

Method	COCO-Stuff		ADE20K	
	MIoU \uparrow	FID \downarrow	MIoU \uparrow	FID \downarrow
Real Images	40.87	-	54.41	-
FreestyleNet [47]	42.42	15.12	44.42	28.45
PLACE [24]	42.23	14.95	60.20	24.51
GLIGEN [22]	-	-	23.78	33.02
ControlNet [56]	27.46	21.33	32.55	33.28
UniControl [29]	-	-	25.44	46.34
Controlnet++ [20]	34.56	19.29	43.64	29.49
Seg2Any (ours)	45.54	19.90	<u>54.46</u>	32.89

Table 3: Ablation results on COCO-Stuff and ADE20K. We conduct ablation experiments on SAA (Semantic Alignment Attention Mask), SSFA (Sparse Shape Feature Adaptation), and CITF (Condition Image Token Filtering).

Methods			COCO-Stuff		ADE20K	
SAA	SSFA	CITF	MIoU \uparrow	FID \downarrow	MIoU \uparrow	FID \downarrow
✓	-	-	43.48	20.57	44.85	33.14
✓	✓	-	45.50	19.06	54.11	32.37
✓	✓	✓	45.54	19.90	54.46	32.89

Table 4: Ablation results on the SACap-Eval benchmark. MIoU* denotes class-agnostic MIoU. The ablation focuses on SAA (Semantic Alignment Attention Mask), SSFA (Sparse Shape Feature Adaptation), AIA \dagger (Training-Free Attribute Isolation Attention Mask) AIA \ddagger (Training-Based Attribute Isolation Attention Mask), and CITF (Condition Image Token Filtering).

Methods					MIoU*	Region-wise Quality				Global-wise Quality			
SAA	SSFA	AIA \dagger	AIA \ddagger	CITF		Spatial \uparrow	Color \uparrow	Shape \uparrow	Texture \uparrow	IR \uparrow	Pick \uparrow	CLIP \uparrow	FID \downarrow
✓	-	-	-	-	89.39	93.36	89.26	85.84	87.86	0.47	21.83	28.09	15.30
✓	✓	-	-	✓	94.16	93.43	89.16	85.61	88.53	0.44	21.81	27.98	15.20
✓	✓	✓	-	✓	94.46	92.37	92.16	87.77	90.23	0.27	21.39	27.79	16.11
✓	✓	-	✓	-	94.94	93.36	90.08	87.40	88.93	0.39	21.57	27.97	17.35
✓	✓	-	✓	✓	94.90	93.89	91.52	88.15	90.12	0.44	21.66	27.87	15.53

5 Conclusion

We propose Seg2Any, a novel segmentation-mask-to-image generation framework that achieves fine-grained layout control by decoupling spatial layout and semantic guidance. Through the integration of sparse entity contours and multi-modal masked attention, Seg2Any simultaneously ensures shape preservation, semantic alignment, and robust attribute control. We further introduce the large-scale SACap-1M dataset and SACap-Eval benchmark to foster open-set S2I research. Extensive experiments validate that Seg2Any achieves new state-of-the-art performance, particularly excelling in generating entities with detailed descriptions.

Limitation. Seg2Any faces resource constraints when generating images with a large number of entities, each accompanied by detailed descriptions. Additionally, our large-scale dataset relies on vision-language models for regional captioning, which inevitably introduces annotation noise that may impact segmentation-mask-to-image generation performance.

6 Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant 62472098).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] 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.

- [3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018.
- [4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024.
- [5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.
- [7] Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*, pages 432–448. Springer, 2024.
- [8] Xueqing Deng, Qihang Yu, Ali Athar, Chenglin Yang, Linjie Yang, Xiaojie Jin, Xiaohui Shen, and Liang-Chieh Chen. Coconut-pancap: Joint panoptic segmentation and grounded captions for fine-grained understanding and generation. *arXiv preprint arXiv:2502.02589*, 2025.
- [9] 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 *International Conference on Machine Learning*, 2024.
- [10] Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4744–4753, 2024.
- [11] Yuchao Gu, Yipin Zhou, Yunfan Ye, Yixin Nie, Licheng Yu, Pingchuan Ma, Kevin Qinghong Lin, and Mike Zheng Shou. Roictrl: Boosting instance control for visual generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23658–23667, 2025.
- [12] 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.
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3, 2022.
- [14] Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7797–7806, 2024.
- [15] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023.
- [16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
- [17] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.
- [18] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [19] Yuseung Lee, Taehoon Yoon, and Minhyuk Sung. Groundit: Grounding diffusion transformers via noisy patch transplantation. *Advances in Neural Information Processing Systems*, 37:58610–58636, 2024.
- [20] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, 2024.

- [21] Yu Li, Xingyu Qiu, Yuqian Fu, Jie Chen, Tianwen Qian, Xu Zheng, Danda Pani Paudel, Yanwei Fu, Xuanjing Huang, Luc Van Gool, et al. Domain-rag: Retrieval-guided compositional image generation for cross-domain few-shot object detection. *arXiv preprint arXiv:2506.05872*, 2025.
- [22] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [24] Zhengyao Lv, Yuxiang Wei, Wangmeng Zuo, and Kwan-Yee K Wong. Place: Adaptive layout-semantic fusion for semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9264–9274, 2024.
- [25] Yuhang Ma, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation. *Advances in Neural Information Processing Systems*, 37:128886–128910, 2024.
- [26] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.
- [27] Marianna Ohanyan, Hayk Manukyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Zero-painter: Training-free layout control for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8764–8774, 2024.
- [28] 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.
- [29] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: a unified diffusion model for controllable visual generation in the wild. In *Advances in Neural Information Processing Systems*, pages 42961–42992, 2023.
- [30] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023.
- [31] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.
- [32] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- [33] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaoje Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.
- [34] 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.
- [35] Large scale Artificial Intelligence Open Network (LAION). Laion-aesthetics-v2. <https://github.com/christophschuhmann/improved-aesthetic-predictor>, 2022.
- [36] Takahiro Shirakawa and Seiichi Uchida. Noisecollage: A layout-aware text-to-image diffusion model based on noise cropping and merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8921–8930, 2024.
- [37] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. 2025.

- [38] Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.14682*, 2025.
- [39] Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025.
- [40] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21096–21106, 2025.
- [41] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableanimator++: Overcoming pose misalignment and face distortion for human image animation. *arXiv preprint arXiv:2507.15064*, 2025.
- [42] Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025.
- [43] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [44] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242, 2024.
- [45] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*, pages 207–224. Springer, 2024.
- [46] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.
- [47] 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.
- [48] Binbin Yang, Yi Luo, Ziliang Chen, Guangrun Wang, Xiaodan Liang, and Liang Lin. Law-diffusion: Complex scene generation by diffusion with layouts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22669–22679, 2023.
- [49] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.
- [50] Zuyao You, Junke Wang, Lingyu Kong, Bo He, and Zuxuan Wu. Pix2cap-coco: Advancing visual comprehension via pixel-level captioning. *arXiv preprint arXiv:2501.13893*, 2025.
- [51] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211, 2024.
- [52] Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John Collomosse, Jason Kuen, and Vishal M Patel. Scenecomposer: Any-level semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22468–22478, 2023.
- [53] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yingda Chen, and Yu Zhang. Eligen: Entity-level controlled image generation with regional attention. *arXiv preprint arXiv:2501.01097*, 2025.
- [54] Hui Zhang, Dexiang Hong, Tingwei Gao, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. 2025.
- [55] Hui Zhang, Dexiang Hong, Maoke Yang, Yutao Cheng, Zhao Zhang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatidesign: A unified multi-conditional diffusion transformer for creative graphic design. *arXiv preprint arXiv:2505.19114*, 2025.

- [56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [57] Shilong Zhang, Peize Sun, Shoufa Chen, Minn Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on regionof-interest, 2024. In *URL <https://openreview.net/forum>*.
- [58] 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.
- [59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017.
- [60] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024.
- [61] Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. 2025.
- [62] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. In *International Conference on Learning Representations*, 2025.
- [63] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis-flux: simple and efficient multi-instance generation with dit rendering. *arXiv preprint arXiv:2501.05131*, 2025.

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: In this paper, we introduce a novel segmentation-mask to image generation approach and describe the contributions in the Abstract and Introduction section (see the Abstract and Introduction).

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: We have discussed the limitations of our work in Section 5.

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: This paper does not include experimental results related to theoretical aspects.

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 detail the experimental setup in Section 4.

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: [No]

Justification: Our data and code are not open-source yet. However, we have provided sufficient details of our method and all the required hyperparameters in the main paper and supplementary material to enable the reproduction. We plan to open-source the data and code in the future.

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: We have specified all the training and test details necessary to understand the results (See Section 4).

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: Due to extensive computation and time cost, we do not report error bars.

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: In this paper, we provide detailed information about the experimental resources, including GPU configurations used in our studies (See Section 4).

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 study presented in this paper conforms to the NeurIPS Code of Ethics.

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: See supplementary material.

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: **[No]**

Justification: This paper does not address issues related to this aspect.

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: All creators and original owners of the assets used in our paper, such as code, data, and models, have been properly credited (See reference). We have explicitly mentioned the licenses and terms of use for each asset and have ensured full compliance with these terms throughout our research.

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: [No]

Justification: The paper has not released new assets yet.

Guidelines: The dataset/code/model include details about training, license, limitations, etc.

- 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 paper does not involve experiments or research related to 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: The paper does not involve crowdsourcing nor research with human subjects.

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: [NA]

Justification: The core method development in this research does not involve LLMs.

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.

A More Details of Condition Image Token Filtering

We analyze the computational cost of Condition Image Token Filtering (CITF) on three benchmark datasets: COCO-Stuff, ADE20K, and SACap-1M. For each dataset, we consider five settings to comprehensively analyze the computational cost: I) without condition tokens, representing the lowest computation. II–IV) CITF applied with the minimum, maximum, and average numbers of retained condition tokens, respectively. V) without CITF, i.e., no token filtering, representing the highest computation. For each configuration, we report the resulting average image generation time and multiply-accumulate operations (MACs).

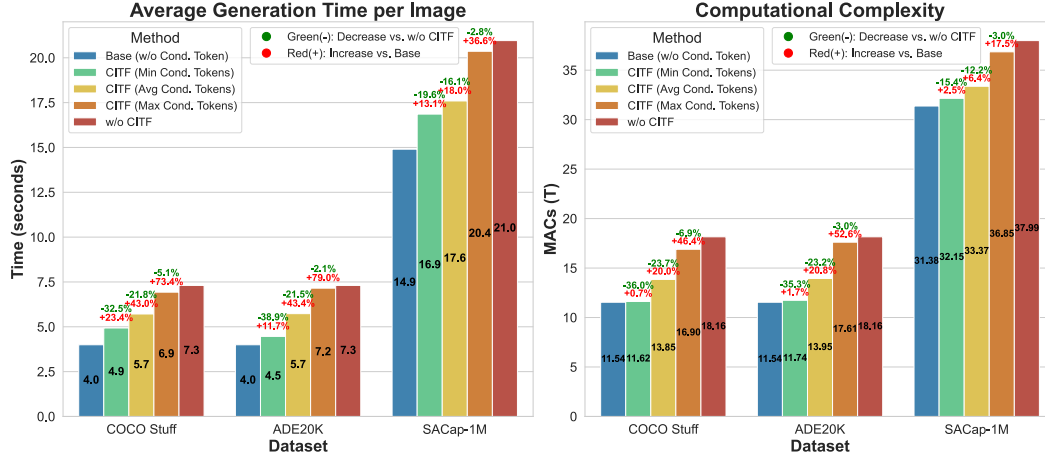


Figure 6: Impact of Condition Image Token Filtering (CITF) on computational cost.

Figure 6 illustrates the results. All experiments are conducted on a single NVIDIA H100 GPU with 32 sampling steps. Each image uses 5 regional text prompts (50 tokens each) and one global prompt (512 tokens). As shown, CITF leads to a notable reduction in both inference time and MACs when the entity contour maps are sparse. When more tokens are retained, computational savings decrease. Overall, CITF provides a simple yet effective mechanism to reduce inference overhead, particularly in cases with sparse entity layouts.



Figure 7: Qualitative results with shape guidance strength ($\gamma = 0.2$), demonstrating flexible scribble-style control.

B More Details of Shape Guidance Strength Modulation

We empirically find that adjusting the strength hyperparameter γ in Shape Guidance Strength Modulation serves as a “free lunch”, enabling scribble-style control without incurring any additional training overhead. In our experiments, we set $\gamma = 0.2$, which provides sufficient flexibility for scribble-style segment masks while still maintaining semantic alignment. The qualitative results are shown in Figure 7.

C More Ablation Study Results

C.1 Ablation on AIA

To further verify the effectiveness of the proposed Attribute Isolation Attention Mask (AIA), we conduct an ablation experiment by removing the AIA module from Seg2Any. As illustrated in Figure 8, the absence of AIA leads to attribute leakage across entities.

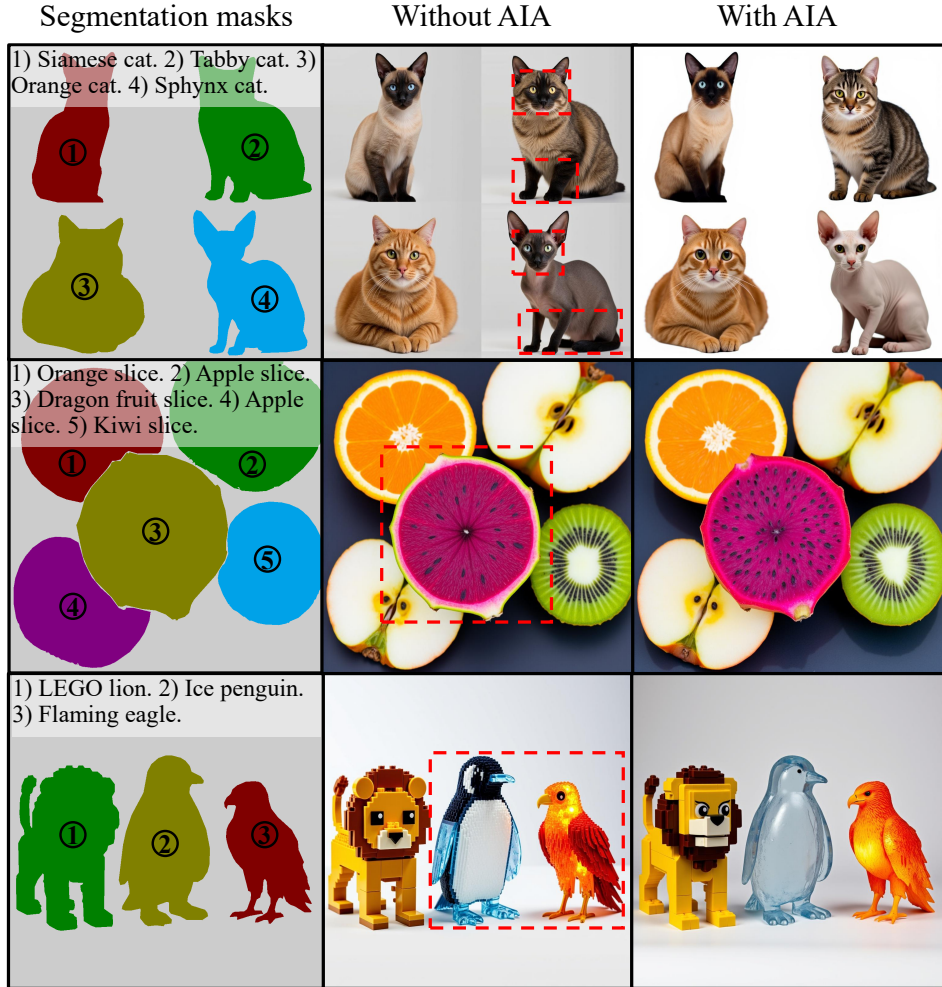


Figure 8: More qualitative ablation comparisons on the Attribute Isolation Attention Mask (AIA). Without AIA, attribute leakage occurs across entities, e.g., fur texture is transferred between cat breeds (row 1), the red peel of dragon fruit is replaced by kiwi’s green peel (row 2), and the LEGO style spreads to other entities (row 3). In contrast, with AIA, such attribute leakage is effectively eliminated.

C.2 Ablation on SSFA vs. PLACE Attention

To ensure a fair comparison between the proposed Sparse Shape Feature Adaptation (SSFA) and PLACE attention [24], we reimplemented PLACE attention [24] under the same FLUX architecture, while keeping all other training setups fixed. Specifically, we use a learning rate of $1e-4$, a batch size of 16, and train for 20k steps on 4 NVIDIA H100 GPUs. As shown in Tables 5 and 6, our model surpasses this PLACE attention variant by (+2.55%) MIoU on ADE20K and (+3.36%) class-agnostic MIoU on SACap-Eval, confirming the superiority of our design in a strictly fair comparison.

Table 5: Comparison with PLACE attention under the FLUX model on COCO-Stuff and ADE20K.

Method	COCO-Stuff		ADE20K	
	MIoU \uparrow	FID \downarrow	MIoU \uparrow	FID \downarrow
SAA	43.48	20.57	44.85	33.14
PLACE attention	44.56	20.10	51.56	32.22
SAA+SSFA	45.50	19.06	54.11	32.37

Table 6: Comparison with PLACE attention under the FLUX model on SACap-Eval.

Method	Class-agnostic MIoU \uparrow	Region-wise Quality				Global-wise Quality			
		Spatial \uparrow	Color \uparrow	Shape \uparrow	Texture \uparrow	IR \uparrow	Pick \uparrow	CLIP \uparrow	FID \downarrow
SAA	89.39	93.36	89.26	85.84	87.86	0.47	21.83	28.09	15.30
PLACE attention	90.8	93.03	89.12	86.36	88.08	0.51	21.90	27.93	15.81
SAA+SSFA	94.16	93.43	89.16	85.61	88.53	0.44	21.81	27.98	15.20

Figure 9 presents the synthesis results on the ADE20K dataset using SAA, PLACE attention, and SAA+SSFA, all trained under the same FLUX architecture. It can be seen that SAA+SSFA generates shapes that best conform to the provided segmentation masks.

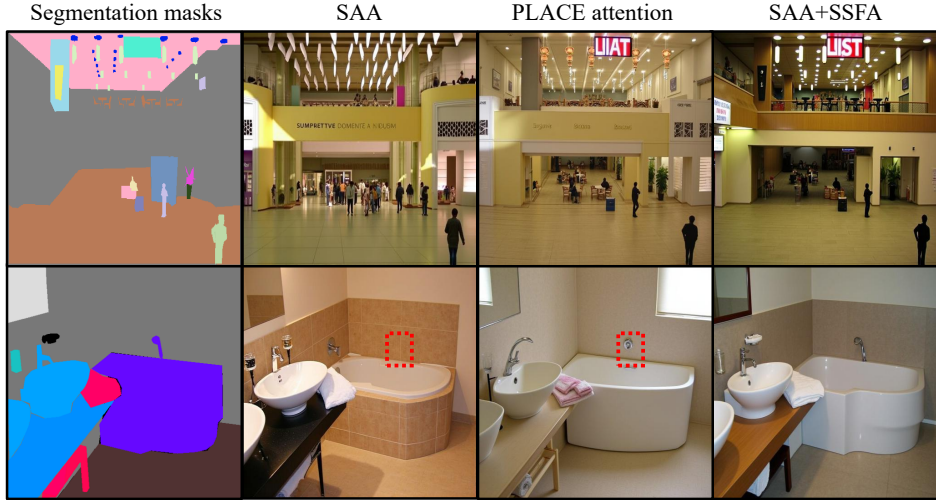


Figure 9: More qualitative ablation comparisons on the Sparse Shape Feature Adaptation (SSFA). All methods are trained under the same FLUX architecture.

D More Qualitative Results

We provide additional qualitative results in Figure 10 to illustrate the effectiveness of our approach. Seg2Any shows strong capability in generating high-quality images that faithfully adhere to detailed entity descriptions, enabling flexible and precise control over complex visual attributes such as color, texture, and shape. Figure 11 presents the qualitative results of Seg2Any in densely crowded scenes.



Figure 10: More qualitative results using the Seg2Any method.



Figure 11: Qualitative results of Seg2Any in densely crowded scenes.

E More Details on Datasets and Benchmarks

E.1 Comparison with Existing Dense Caption Datasets

Table 7 presents a detailed comparison between our SACap-1M dataset and several recent dense caption datasets. Unlike previous datasets such as PixelLM-MUSE [33], Osprey [51], COCONut-PanCap [8] and Pix2Cap-COCO [50] which are constructed upon closed-set label spaces (e.g., LVIS [12] and COCO [23]), our proposed SACap-1M dataset provides open-set segmented entities, enabling much broader generalization and flexibility in open-set segmentation-mask-to-image generation. Compared to GLaMM-GranD [31], also derived from SA-1B [16], our SACap-1M achieves significantly higher segmentation mask density and caption granularity by employing the advanced open-source Qwen2-VL-72B [43] vision-language model to generate more accurate and fine-grained regional captions.

Table 7: Comparison of dense caption datasets. Note that "Avg. Words" indicates the word count per regional caption, and "Avg. Masks" denotes the average number of masks per image.

Dataset Name	Image Source	Image Number	Annotated by	Avg. Words	Avg. Masks
PixelLM-MUSE [33]	LVIS [12]	246K	GPT-4V [1]	11.3	3.7
Osprey [51]	COCO [23] PACO-LVIS [30]	724K	GPT-4V [1]	38.7	-
GLaMM-GranD [31]	SA-1B [16]	11M	GPT4RoI [57] GRiT [45]	5.8	4.4
COCONut-PanCap [8]	COCO [23]	118K	GPT-4V [1] and human correction	16.6	13.2
Pix2Cap-COCO [50]	COCO [23]	20k	GPT-4V [1] and human correction	22.94	8.14
SACap-1M (ours)	SA-1B [16]	1M	Qwen2-VL-72B [43]	14.1	5.9

E.2 Data annotation pipeline

We construct our dataset from the SA-1B [16] dataset through a multi-stage filtering and annotation process:

I) **Image Filtering.** Images are filtered based on size and aspect ratio, keeping those with width and height between 1000 and 3000 pixels and aspect ratios from 0.6 to 1.8. To further ensure visual quality, images with a LAION-Aesthetics [35] score below 5 are excluded.

II) **Entity Extraction.** While SA-1B provides numerous accurate, class-agnostic segmentation masks—averaging over 100 per image—many are nested and redundant. To mitigate this redundancy,

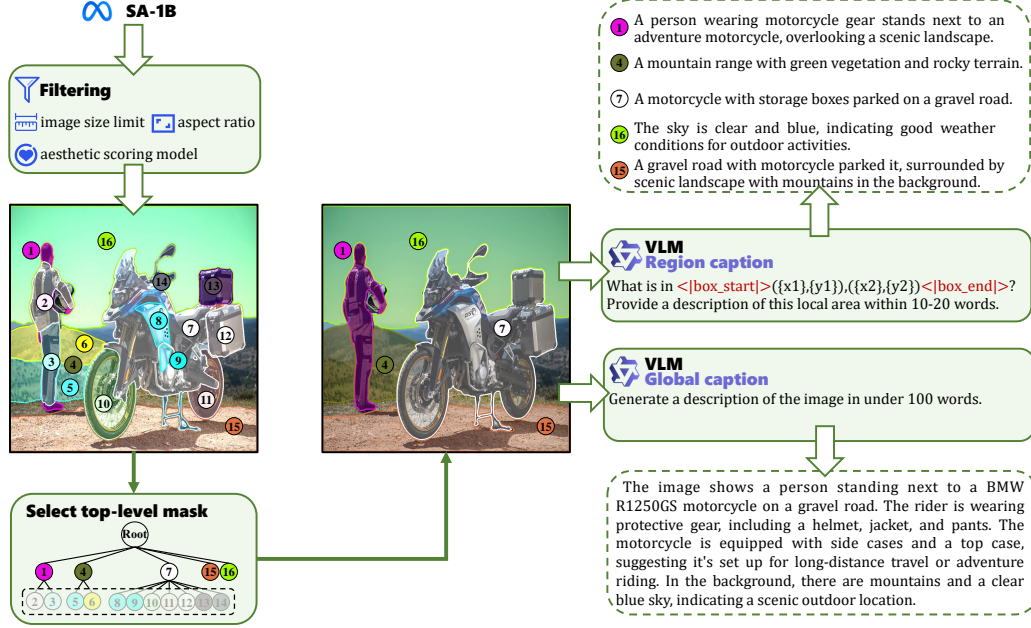


Figure 12: An overview of the data annotation pipeline.

we retain only the top-level masks, which generally correspond to the primary objects in the image, and discard any masks fully enclosed by others (as shown in Figure 12). Additionally, masks occupying less than 1% of the image area are considered too small and removed. Finally, only images containing between 1 and 20 valid masks are retained for further annotation.

III) Regional and Global Caption Annotation. We utilize the vision-language model Qwen2-VL-72B [43] for both regional and global captioning. For regional captioning, we incorporate the bounding box coordinates of each segment into the text prompt (see Figure 12), enabling the model to produce accurate and context-aware descriptions of each local region. This approach yields detailed annotations, with an average of 14.1 words per entity and 58.6 words per image, supporting fine-grained segmentation-to-image generation.

Finally, we present SACap-1M, a large-scale layout dataset containing 1 million image-text pairs and approximately 5.9 million segmented entities. As illustrated in Figure 13, most images contain a moderate number of masks, while images with a large number of masks constitute the long tail of the distribution. Figure 14 shows some examples from SACap-1M.

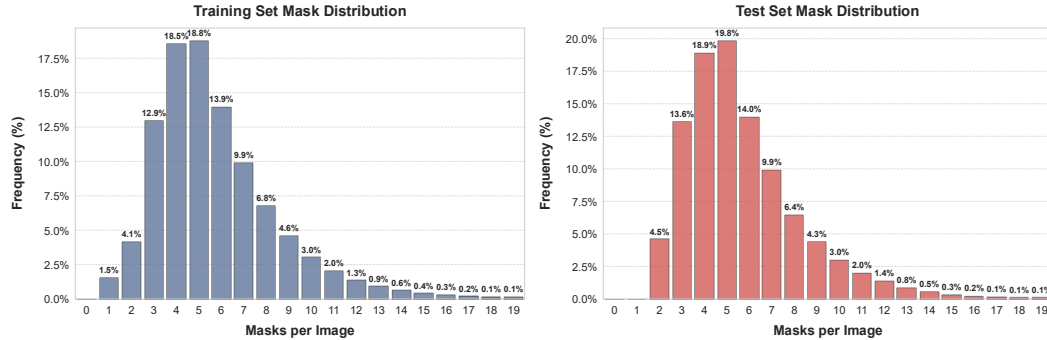


Figure 13: The distribution of the number of segmentation masks per image across the training and test sets.



Figure 14: Examples from the SACap-1M dataset.

E.3 SACap-Eval Benchmark

We construct SACap-Eval, a benchmark derived from SACap-1M, designed to assess the quality of segmentation-mask-to-image generation. The benchmark comprises 4,000 samples, with an average of 5.7 entities per image. Evaluation is conducted from two perspectives: Spatial and Attribute. Both aspects are assessed using the vision-language model Qwen2-VL-72B [43] via a visual question answering manner.

Spatial Score. For each segmentation mask, we first crop the corresponding region from the image and then prompt the VLM to determine whether the target entity is located within this area, allowing responses of either “Yes” or “No”. The spatial score is obtained by computing the ratio of “Yes” answers to the total number of entities.

Attribute Score. To compute the attribute score, the region specified by each segmentation mask is cropped from the image, after which the VLM determines whether the entity inside this area satisfies the described attributes. Each attribute (*e.g.* color, shape, or texture) is examined separately using visual question answering, and the score is calculated in the same manner as the spatial score.

F Broader Impacts

The proposed method for segmentation-mask-to-image synthesis has potential applications in controllable image generation [21, 14], video generation [40, 41, 39], and the construction of image segmentation datasets. However, like other generative models, its misuse could result in the production of misleading or inappropriate content. Our approach may also inherit biases present in the training datasets, potentially reinforcing certain stereotypes. Responsible usage and further investigation into fairness and robustness are important, but a comprehensive analysis is beyond the scope of this work.