A More Implementation Details

A.1 Base Model Selection

In our framework, different diffusion backbone models exhibit variations under fine-tuning-free settings for achieving high-fidelity multi-instance generation. We evaluated three variants of the FLUX family of models as potential backbones: FLUX.1-Dev (a general image generation model) (Labs, 2024b), FLUX.1-Fill (a local inpainting model) (Labs, 2024a), and FLUX.1-Kontext (an editing model) (Labs et al., 2025). While existing multi-subject-driven generation methods without layout control (Wu et al., 2025c; Hu et al., 2025) without attention masks have successfully utilized FLUX.1-Dev, our experiments showed a significant limitation: without additional fine-tuning, FLUX.1-Dev failed to produce coherent images when an attention mask was applied. In contrast, both FLUX.1-Fill and FLUX.1-Kontext demonstrated the ability to generate images correctly with the attention mask. Among these two, FLUX.1-Kontext exhibited a noticeably superior capacity for identity preservation. Therefore, we chose FLUX.1-Kontext as the backbone, as it fits better with our framework's attention-masking strategies and identity preservation goals. To isolate the impact of the backbone, we re-implemented a baseline method MS-Diffusion on FLUX.1-Kontext backbone for ablation studies in Appx. B.7. And we conducted qualitative comparisons with other state-of-the-art methods built upon equivalent FLUX-Family backbones in Appx. D.1.

A.2 Details of Compositing Layout Image

Our Contextual Layout Anchoring (CLA) mechanism relies on a meticulously constructed composite layout image to achieve robust spatial control. The design of this mechanism is conceptually inspired by the principles of multi-feature learning (Yang et al., 2013), which suggests that combining evidence from heterogeneous feature representations can lead to more robust semantic understanding. This process involves two key steps: determining the optimal composition order for all instances and then precisely placing each instance onto the canvas.

A correct composition order is crucial for multi-instance synthesis, especially when handling occlusions and complex overlaps. We propose a dynamic sorting algorithm, **Instance Layering Prioritization**, which first handles explicit containment relationships by prioritizing instances whose masks are completely contained within another's. For all other candidate instances, we use a **hybrid priority scoring system** to simulate the natural layering of objects. We utilize a pre-processing step to obtain each instance's precise effective area (Ravi et al., 2024). The priority score P_i is calculated as:

$$P_i = \alpha \cdot \mathcal{A}(\text{instance}_i) + \beta \cdot \left(1 - \sum_{j \neq i} \text{IoU}(\text{instance}_i, \text{instance}_j)\right) + \lambda \cdot \text{RandomFactor},$$

where $\mathcal{A}(\text{instance}_i)$ is the area of instance i, $\text{IoU}(\text{instance}_i, \text{instance}_j)$ is the Intersection over Union between instances, and α, β, λ are hyperparameters.

The proposed hybrid priority scoring system is designed to simulate general, high-probability composition orders for model training. The introduction of the random factor enhances data diversity and model robustness during training. Despite the Identity Consistency Anchoring (ICA) mechanism, the overlap relationships can still influence the final generated image. Thus, for inference, a user-provided layout offers a more direct and customized form of control.

B More Details and Analysis on Experiments

B.1 Experimental Details on LAMICBench++

As shown in Tab. 1, we benchmark our method against several strong baselines, including single-image-editing models (Wu et al., 2025a; Labs et al., 2025) and closed-source commercial models (OpenAI, 2024; DeepMind, 2025). Since these two categories of models require different input modalities and instruction methods, we prepared distinct inputs for a fair evaluation:

Table 5: **Ablation study on different position indexing strategies.**

Position Indexing Strategies	ITC	AES	IDS	IPS	AVG
w/o x and y offsets	90.53	57.92	21.54	74.57	61.14
w/o y offsets	90.22	57.60	21.01	74.02	60.71
w/o x offsets	**91.53**	57.87	26.71	74.59	62.67
w/o editing token	90.95	<u>58.11</u>	<u>32.64</u>	<u>75.28</u>	<u>64.25</u>
w/ offsets and editing token	<u>91.38</u>	**58.24**	**32.72**	**76.32**	**64.66**

Single-Image-Editing Models These models are primarily designed for single-image editing, meaning they must process all instances combined within a single input image. We thus provided our manually composited layout images, ensuring minimal overlap across instances to avoid ambiguity. The following prompt template was used: "*Use the objects or humans in the image to create a new image that shows* '{PROMPT}'. *Preserve object features and human identities (if any, including facial details). You may fill in the background with appropriate details to achieve a natural and aesthetically harmonious result.*"

Closed-Source Commercial Models These are general-purpose multi-modal models that accept multiple input images. To guide them to perform the specific multi-instance generation task while preserving identities, we relied on a dedicated prompt alongside the references. The prompt template was: "*Generate a high-quality image of* '{PROMPT}'. *Use the provided references, preserve object features and human identities (if any, including facial details).*"

B.2 Ablation on Position Indexing

As specified in Sec. 3.2, we extend the position indexing strategy from prior work (Wu et al., 2025c), which uses non-overlapping indices for explicit spatial delineation of objects, crucial for multi-instance distinction. The ablation results on LAMICBench++ are presented in Tab. 5. Our findings validate the importance of the positional encoding components:

- **The Necessity of Explicit Spatial Separation**: The results strongly confirm the importance of explicit spatial separation achieved via x and y offsets. Removing any offset component (e.g., *w/o x and y offsets*) significantly degrades all metrics, confirming the necessity of these offsets for the non-overlapping indexing mechanism. Its superiority is evidenced by the large margin observed in the IDS, suggesting better token distinction across different reference images in token sequence.
- **Role of Editing Token**: Our complete strategy includes an "editing token" (setting the first component of the position index triplet to 1). The ablation *w/o editing token* (setting it to 0) shows minimal performance difference in the final metrics. However, omitting this token actually resulted in a certain degree of gradient instability during the initial phase of training, which justifies its inclusion for a more stable optimization process.

B.3 Details of DPO Fine-tuning

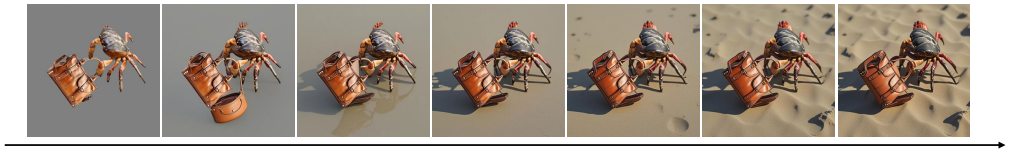
B.3.1 Qualitative Results of DPO Fine-tuning Process

The visualization in Fig. 7 illustrates the efficacy of Direct Preference Optimization (DPO) in enhancing image generation, particularly by mitigating the issue of rigidly copying the layout image with blank backgrounds. To demonstrate this, we intentionally select an input and seed configuration that typically results in a minimal background scene.

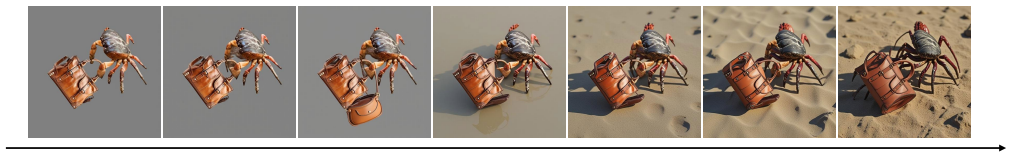
The initial result on the left of Fig. 7a correctly anchors the main subject but suffers from the blank background issue, failing to render any environmental context. As the fine-tuning process advances, the model begins to introduce more naturalistic details, first by generating a realistic shadow and subsequently by adding a simple yet coherent background. Upon convergence, the final image on the right features a rich, detailed background, effectively demonstrating DPO's ability to enrich the overall scene while strictly preserving the subject's layout.

Fig. 7b shows how the DPO β parameter affects the generation quality. A high β value may limit the model's capacity for meaningful tuning, whereas an overly low β value can cause the model to

follow the preference data too aggressively, risking a loss of the subject's identity during convergence (as seen with the leather bag when $\beta = 50$). Our results validate that when β is correctly calibrated, DPO substantially improves image quality with minimal compromise to the subject's identity.



(a) Results during the DPO fine-tuning process ($\beta = 100, \text{Rank} = 256$).



(b) Results for the same input and fixed seed of DPO fine-tuning ($\text{Rank} = 256$).

Figure 7: **Results for the same input and fixed seed of DPO fine-tuning.**

B.3.2 Image Quality Assessment on DPO Fine-tuning

In order to fully assess the impact by DPO fine-tuning on image quality, we report the quality assessment results in Tab. 6. However, the FID score on LayoutSAM-Eval may not be highly referential, not only because FID is less sensitive to fine-grained image details, but also because the calculation process necessitates severe distortion by resizing the structurally diverse ground-truth images of this benchmark, thus failing to reflect the objective quality of the generated outputs. As shown in Fig. 8, the model with the best FID score on LayoutSAM-Eval exhibits a less satisfactory overall visual effect, specifically in background richness, light and shadow rendering, and fine detail processing, compared to the results obtained using a smaller DPO β value.

To better assess the overall visual quality and fidelity of the generated images, despite the metrics AES and FID score from the benchmarks, we conducted a **user study**. In this user study, we sampled 20 sets of inputs and seeds from each of the two benchmarks (LAMICBench++ and LayoutSAM-Eval). For each sampled set, we generated results using 6 different DPO β's, alongside a result from the model without DPO fine-tuning. This yielded a total of 7 images for comparison per set. 10 reviewers were asked to select **two** images with the best quality and **two** images with the worst quality in each group of images. Each "best quality" selection added one point to an image's score, while each "worst quality" selection deducted one point. The results are normalized and recorded at *User Preference* metric in Tab. 6. The comparison results further demonstrate that DPO fine-tuning will enhance image quality, and a smaller β (a more aggressive DPO) generally leads to better quality.

Table 6: **Image quality assessment results on benchmarks and user study.**

DPO β	AES	FID	User Preference
100	**57.97**	55.97	**0.54**
250	57.58	55.60	0.37
500	57.57	**55.32**	0.16
750	57.22	55.53	0.03
1000	57.10	55.65	-0.11
1500	56.69	55.70	-0.43
w/o DPO	54.19	55.93	-0.56

B.3.3 Qualitative Validation on Final Model

Following the ablation study in Tab. 4, we applied the optimal DPO setting ($\beta = 1000$) to our final ContextGen model, which utilizes LoRA Rank 512. The qualitative results presented in Fig. 9

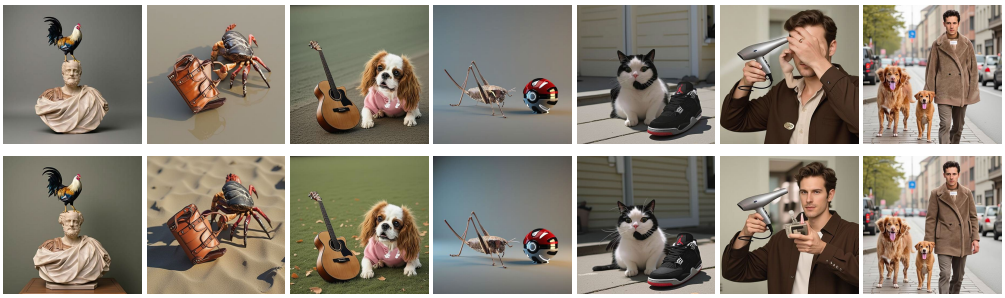


Figure 8: **Qualitative Comparison: DPO $\beta = 500$ (FID-Optimal) vs. DPO $\beta = 100$.** The comparison showcases the model fine-tuned with the FID-optimal setting (DPO $\beta = 500$, upper panel) versus a more aggressive setting ($\beta = 100$, lower panel). All corresponding images were generated with identical inputs and random seeds. The columns are specifically grouped to demonstrate the quality differences: Columns 1–3 highlight background richness; Columns 4–5 illustrate light-and-shadow rendering; and the final two samples reveal differences in fine identity details. Zoom for more details.

Figure 9: **Qualitative comparison of final Model** ($\mathrm{Rank} = 512$)**: before vs. after DPO Fine-tuning.** The upper panel shows images generated before DPO fine-tuning, and the lower panel shows the results after. All corresponding images were produced using the same random seed.

strongly validate that DPO fine-tuning significantly enhances the visual quality and overall fidelity of the outputs from this high-rank configuration. This enhancement is particularly evident in several key aspects: increased richness in background details, improved flexibility in instance composition and more natural rendering of light and shadow.

B.4 Results and Analysis on Model's Generation Flexibility

The model's final output represents a crucial trade-off between Fidelity (adherence to reference image identities and given layout) and Flexibility (the ability to incorporate text descriptions, especially for inter-subject interactions). The richness of input text prompt significantly modulates this balance. When a simple or minimal prompt is provided, the model inherently prioritizes high fidelity, leaning towards preserving the exact details and posture defined by the reference images. Conversely, by enriching the prompt with complex descriptions and detailed interactions between subjects, the model will exhibit greater flexibility in the generation results.

Fig. 10 demonstrates how prompt complexity influences the generation outcome while the layout image remains constant. For instance, in the first panel of Fig. 10, when the action is described as "having a conversation", the model exhibits flexibility by removing the sword from the pixelated figure, a detail conflicting with the context. Furthermore, when the prompt is "in a fierce dual", the figure's posture is greatly modified to align with the dynamic description.

B.5 ICA: An Indispensable Component Handling Identity in Overlapped Regions

Although a pure CLA method (using only a layout image as contextual input) exhibits scores similar to strategies with ICA in the aggregated results of Tab. 3, the critical necessity of the ICA component

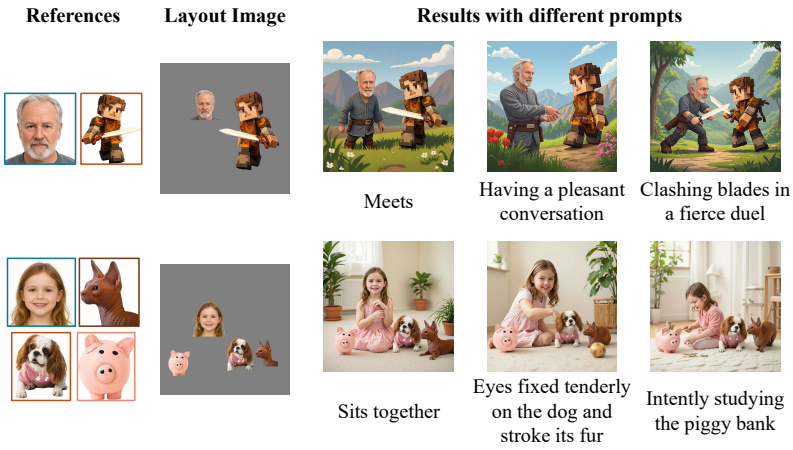


Figure 10: **Flexible compositions with different prompts.**

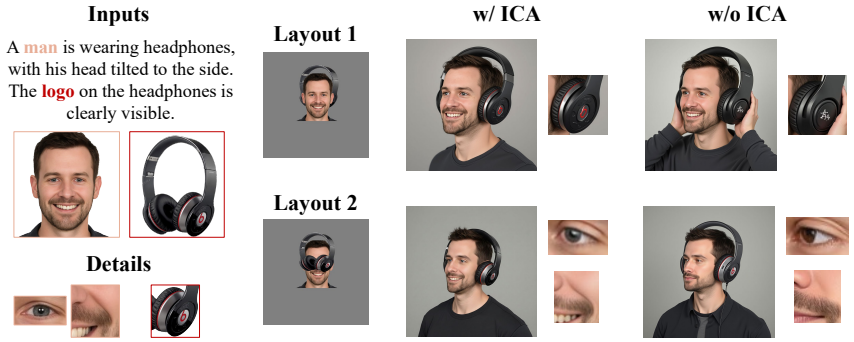


Figure 11: **Results of overlapping layout on methods with and without ICA.**

is explicitly demonstrated when handling layouts with overlaps. For generating layout involving overlapping instances, ICA is critical because the reference images in the token sequence provide extra identity information for the regions being occluded by other instances, mitigating identity loss in these overlapping areas.

Fig. 11 illustrates this effect across two scenarios: For Layout 1, subtle details like the red logo of the headphones are successfully preserved by the method with ICA but completely abandoned by the method without ICA. Conversely, in the occluded scene of Layout 2, the ICA method better retains fine-grained facial identity features—such as the color of the eyes and the presence of the nasolabial fold—compared to the non-ICA method, which loses these crucial details.

However, these overlapping cases are relatively rare in the current benchmark, and the differences on the image lead to a limited overall impact on the macro-level evaluation indicators. But these qualitative results demonstrate that ICA is an indispensable component for ensuring identity preservation in such cases. And its absence significantly impacts the final image's visual quality and detailed fidelity.

B.6 Analysis on Layer-wise ICA Ablation

As shown in Tab. 3, applying ICA to different layers (grouped by First-19, Mid-19 and Back-19) reveals an interesting result. Our ablation study revealed a gap between two distinct sets of strategies: the first set, comprising $\{F + M + B, F + M, B\}$, consistently yielded lower scores, while the second set, $\{F, \text{w/o CLA}, M + B, M\}$, achieved closely clustered higher scores. This result suggests the existence of intrinsic properties within the FLUX-DiT layers, indicating potential functional difference where performance is sensitive to the placement of attention mechanism.

- **F-19: Strategy to Minimize Modality Interference.** Prior work CreatiLayout (Zhang et al., 2024) extending the F-19 blocks has shown a powerful control in both spatial and attribute. And as

stated this work, no further modification is made to the other blocks (Single Stream Blocks, i.e., Mid-19 and Back-19 blocks) in order to reduce the potential interference among modalities (i.e., text, layout, and image). It does make sense in considering the Front-19 blocks contains a huge amount of parameters compared to the others. Adding ICA to these blocks is sufficient, while applying ICA to more blocks lead to some side effect like identity distortions. This aligns with our observation that $F + M + B$ and $F + B$ strategies underperform others.

- **M-19: Critical Role in Attribute Binding.** Prior work DreamRenderer (Zhou et al., 2025b) demonstrates through ablation study that Mid-19 blocks play a relatively more critical role in attribute binding. Our ablation study validates the finding, for applying ICA to Mid-19 blocks yields the best performance in our task on benchmark score.
- **F-19 Protection and Similar Performance.** The four strategies, F, M, $M + B$, and no ICA, yield relatively close performance scores because they all avoid posing further disruptions to the Front-19 blocks. And we can derive from this result that these four strategies do not have a very significant difference in overall performance.
- **B-19: Detrimental Impact on Final Refinement.** Conversely, applying ICA solely to the Back-19 blocks proves detrimental to performance, which is also aligned with the ablation study in DreamRenderer. For the Back-19 blocks serve as final and sensitive post-processing layers in each noise prediction step. Introducing attention mask exclusively at this late stage without any modifications to the earlier blocks will affect the refinement process, leading to problems like artifacts and blurring identity.

We choose Mid-19 from the last four strategies with very close score for the following reasons:

- **Performance**: Applying ICA to the Mid-19 blocks demonstrates an advantage on LAMICBench++ average score over other strategies.
- **Indispensability** of ICA: The analysis in Appx. B.5 has revealed the indispensability of ICA.
- **Efficiency**: Considering that the computational cost of applying the attention mechanism scales with both the number of blocks and the block capacity (such as F-19 blocks with much denser parameters), we selected the Mid-19 configuration for a relatively efficient approach to achieve better performance.

B.7 MS-FLUX: A Controlled Case Study of Strategy Failure on Strong Backbone

Previous image generation methods based on older diffusion models (e.g., Stable Diffusion) have shown considerable improvements over their base architectures. For instance, MS-Diffusion (Wang et al., 2025), built on SDXL, achieved image-guided layout control via its ***Grounding Resampler*** mechanism based on SD-Unet cross attention architecture. This Resampler fuses reference image content with explicit spatial and phrase information into multi-modal prompt tokens to guide generation.

To quantify the contribution of our ContextGen mechanisms and assess their synergistic fit with the FLUX backbone, we implement a competitor **MS-FLUX**. This implementation adapts a similar *Grounding Resampler* architecture onto the FLUX-DiT. We fine-tuned MS-FLUX with all the *Grounding Resampler*'s parameters trained and a LoRA Adapter (R=512) applied to the DiT blocks. After 50K training steps on IMIG-100K, we evaluate this model on COCO-MIG and LAMICBench++ for a comparison, the results of which are presented in Tab. 7. This result demonstrates **severely poor identity preservation** coupled with a **complete degradation of layout control** of MS-FLUX. The method of integrating reference images, phrases and layout information together as a text-prompt-like token sequence, which works on SD-Unet, works poorly on FLUX-DiT. In summary, this confirms that despite the strong contextual capability of FLUX.1-Kontext (FLUX-DiT) over old backbones, not all methods (even those that seem reasonable in principle or already worked on previous backbones, just like this *Grounding Resampler* in MS-Diffusion) will work on FLUX backbone. This further suggests that our specialized mechanisms provide a more correct design for achieving better layout control and identity-consistency on the FLUX backbone.

To ensure complete transparency regarding our comparative experiments and to fully document the controlled competitor, we include the implementation code details of MS-FLUX below. The code

Table 7: **Quantitative result of MS-FLUX on key metircs of LAMICBench++ and COCO-MIG.**

Method	SR	I-SR	mIoU	IDS	IPS
MS-FLUX	0.38	7.3	18.71	3.20	60.22
MS-Diffusion	4.50	28.22	34.69	9.06	69.75
Ours	33.12	69.72	65.12	32.72	76.32

involves three core stages: Resampler initialization, multimodal input encoding, and feature injection into the Transformer.

- **Resampler Initialization and Input Preparation.** The `Grounding Resampler` is initialized within custom `LightningModule` to match the dimensions of the FLUX Transformer. During the training step, image, layout, and phrase information are prepared.

```
# Initialization (in the LightningModule's setup method)
from .projection import Resampler
# ...
self.grounding_resampler = Resampler(
   dim=1280,
   depth=4,
   dim_head=64,
   heads=20,
   num_queries=512,
   embedding_dim=self.flux_pipe.transformer.config.in_channels,
   output_dim=self.flux_pipe.transformer.context_embedder.out_features,
   ff_mult=4,
   latent_init_mode="grounding",
   phrase_embeddings_dim=self.flux_pipe.text_encoder.config.projection_dim,
).to(self.target_dtype)
self.grounding_resampler.requires_grad_(True).train()
```

- **Encoding Reference and Spatial Information.** The `prepare_reference_latents` function encodes the reference images into latents. Concurrently, instance phrases and bounding boxes are encoded and assembled as "Grounding Keywords".

```
# Preparing inputs within the Custom Lightning Model's step() method
# ...
# 1. Extract and normalize boxes
phrases = [inst["phrase"] for inst in instance_info[0]]
boxes = [inst["bbox"] for inst in instance_info[0]]
boxes = torch.stack(boxes).to(self.device, self.target_dtype)
boxes = boxes / torch.tensor([width, height, width, height],
    device=self.device, dtype=self.target_dtype)

# 2. Encode Instance Phrases using FLUX's text encoder
phrase_input_ids = []
for phrase in phrases:
   # ... (Tokenizer call)
   phrase_input_ids.append(phrase_input_id)
phrase_input_ids = torch.stack(phrase_input_ids)
phrase_input_ids = phrase_input_ids.view(-1, phrase_input_ids.shape[-1])
phrase_embeds =
    self.flux_pipe.text_encoder(phrase_input_ids.to(self.device)).pooler_output

# 3. Assemble Grounding Keywords
grounding_kwargs = {
   "boxes": boxes,
   "phrase_embeds": phrase_embeds,
   "drop_grounding_tokens": [0],
}

# 4. Prepare reference image latents
```

```
reference_info_dict = prepare_reference_latents(
    # ... (calling prepare_reference_latents encodes the reference
        images into latents)
)
instance_latents = reference_info_dict["instance_latents"][0]
instance_latents = torch.cat(instance_latents, dim=0)
# ...
```

- **Resampler Processing and Feature Injection.** The `instance_latents` (visual features) and `grounding_kwargs` (spatial and phrase features) are passed to the Resampler. The Resampler's output is then concatenated into the token sequence input of the FLUX Transformer layer.

```
# Calling the Grounding Resampler
img_prompt_hidden_states = self.grounding_resampler(
    x=instance_latents,
    grounding_kwargs=grounding_kwargs,
)
img_prompt_hidden_states = img_prompt_hidden_states.reshape(bs, -1,
     img_prompt_hidden_states.shape[-1])

# Injecting the Resampler output into the custom Transformer forward
    pass
transformer_out = FluxTransformer2DModel_forward(
    self=self.flux_pipe.transformer,
    hidden_states=x_t,
    # ... (standard FLUX arguments)
    img_prompt_hidden_states=img_prompt_hidden_states, # <-- Resampler
        Output Injection
    txt_ids=text_ids,
    img_ids=img_ids,
    img_prompt_ids=img_prompt_ids,
    joint_attention_kwargs=None,
    return_dict=False,
)
```

- **Context Aggregation.** In `FluxTransformer2DModel_forward`, the Resampler's output (`img_prompt_hidden_states`) represents the fused image, bounding box, and phrase information. This feature vector is concatenated with the standard text (`encoder_hidden_states`) to form a unified, comprehensive context for the Transformer.

```
# Context Aggregation Logic (inside the custom
    FluxTransformer2DModel_forward)

# 1. Process standard text tokens
encoder_hidden_states = self.context_embedder(encoder_hidden_states)

# 2. Integrate the MS-FLUX Resampler Output
encoder_hidden_states = torch.cat((encoder_hidden_states,
    img_prompt_hidden_states), dim=1)

# 3. Update corresponding position indices
ids = torch.cat((txt_ids, img_prompt_ids, img_ids), dim=0)
image_rotary_emb = self.pos_embed(ids)

# ... (Encoder hidden states are then passed to the double transformer
    blocks)
```

- **Inference Time Injection.** During inference, the pre-calculated Resampler output is consistently passed into the Transformer at every denoising step.

```
# Inference logic in the custom pipeline forward function
# ...
noise_pred = (
```

```
    # ... (Standard Transformer call if reference_dict is None)
    if reference_dict is None
    else FluxTransformer2DModel_forward(
        self=self.transformer,
        hidden_states=latent_model_input,
        img_prompt_hidden_states=reference_dict["img_prompt_hidden_states"],
            # <-- Inference Injection
        timestep=timestep / 1000,
        # ... (other arguments)
        return_dict=False,
    )[0]
)
# ...
```

C More Details on Multi-Instance Dataset

C.1 Brief Survey and Discussion on Existing Multi-Instance Datasets

Various approaches have been adopted to construct multi-instance datasets. Prior works like OmniGen series (Wu et al., 2025b; Xiao et al., 2024b) and MS-Diffusion (Wang et al., 2025) use a cross-verification method to match and pair group images and corresponding individual images from web images or video frames. And some methods begin to leverage synthetic data for training. For example, UNO (Wu et al., 2025c) employs a co-evolutionary data generation and training process using the model itself, and XVerse (Chen et al., 2025a) add data generated by FLUX.1-Dev as supplementary to high-aesthetic-quality images.

However, the existing datasets contain some limitations as robust training resources for spatial-aware multi-instance generation.

- **Insufficient Instance Count and Lack of Layout Annotations.** Though datasets from general-purposed generation methods or these above mentioned subject-driven works often exhibit a relatively high overall quality, identity consistency and a large overall scale, the data with more than 3 subjects is rather limited. Critically, none of these datasets provide detailed spatial or layout annotations.
- **Restricted Subject Variation between Reference and Target.** MS-Diffusion's dataset, while featuring more subjects per image and layout annotations, has a crucial limitation mentioned in the original paper: the low identity matching rate during cross-verification process across the video frames often necessitates using parts of the ground-truth image directly as reference image, which will limit the necessary variations and meaningful transform between the subjects in reference and target.

To address these limitations, we propose the pipeline in Sec. 3.3 to construct our own IMIG-100K dataset. We choose to use the synthetic data for the following reasons:

- **Real-World Data Scaling Challenges.** The collection and matching process in real-world data is relatively less flexible and more difficult to scale in both diversity and quantity.
- **Successful Precedent of Synthetic Data Usage.** The successful practice in previous work, such as UNO, has demonstrated the effectiveness of evolving synthetic data in training robust subject-driven generation models.

C.2 More Details of IMIG-100K Dataset

Sec. 3.3 provides a general introduction to the synthesis and architectures of our dataset. More details and samples of our IMIG-100K Dataset are presented as follows.

Data Diversity We have designed multiple prompt generation templates to include more scenes and styles. These templates are randomly combined and fed to LLMs maximize the diversity of the resulting image generation prompts. We also use different LLMs (i.e. DeepSeek v3 (DeepSeek-AI, 2025), Gemini 2.5 Flash (DeepMind, 2025) and GPT-4o (OpenAI, 2024)) for lexical and stylistic

Figure 12: **Samples with multiple styles in IMIG Dataset.**

variations. Besides the realistic style (natural style), as illustrated in Fig. 12, we also include extra styles like fantasy and anime, which account for more user scenarios. The proportion of natural scenes remains high, accounting for approximately 85% to 90% of the total data.

Detailed Spatial Annotations Every instance's is meticulously annotated with detailed spatial information, including the instance phrase, its valid area's bounding box and mask, and its corresponding position within the ground-truth image. The third subset of IMIG-100K—the flexible composite part—features more detailed annotations. Recognizing that many user scenarios only provide a facial avatar instead of a half or full-body image, we use face-related tools (Xin, 2022; Guo et al., 2021) to restore and standardize the faces in reference images into aligned facial images, as shown in Fig. 14. Face pairs in the ground-truth images (composited images) and corresponding reference images are strictly filtered by face quality and similarity. Crucially, the positions of valid faces are annotated (indicated by the green round bounding boxes in Fig. 14) across composited images, corresponding reference images and aligned face images. This enables a precise position and scale alignment when positioning and sizing reference instances in canvas in training process, instead of relying on a very rough bounding box of the entire instance.

Overall Quality Our dataset has undergone strict quality assessment (Kirstain et al., 2023; Boutros et al., 2023) for both reference images and ground-truth images. And we leverage MLLMs to ensure the semantic consistency between reference images and generated captions. This aligns with recent efforts in multi-context understanding (Xu et al., 2025) and multimodal condition alignment (Zhou et al., 2025a), which both emphasize maintaining logical and semantic integrity across complex visual inputs. Additional samples illustrating the quality and annotation detail are provided in Figs. 13 and 14.

D More Results and Demos

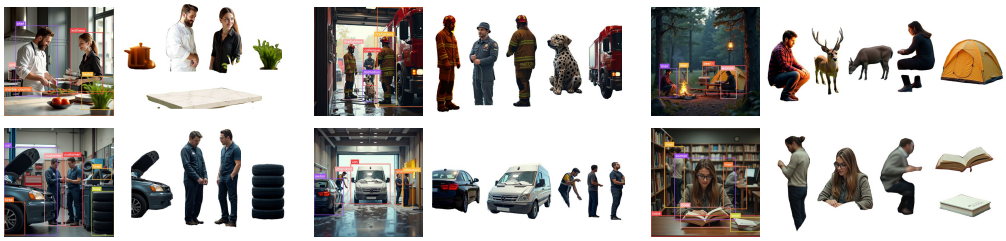
D.1 Qualitative Results Comparing across FLUX-based Methods

Several FLUX-based methods are included in our comparison on LAMICBench++: vanilla FLUX.1-Kontext, LAMIC (based on FLUX.1-Kontext), UNO and DreamO (both based on FLUX.1-Dev). Given that these competitors and our method base on highly similar (or identical) foundational backbones, we provide additional qualitative results in Figs. 15 and 16 for a direct and fair visual comparison.

D.2 More Results on COCO-MIG

Full quantitative result on COCO-MIG benchmark is shown in Tab. 8. Our method establishes a new state-of-the-art on all key metrics, achieving the highest average Success Rate (33.12%) and mIoU (65.12%) among all compared methods. Notably, our performance advantage is most pronounced in complex, high-instance-count scenarios. For L_4, L_5, and L_6 levels, our method significantly outperforms all baselines with a Success Rate of **28.12%**, **23.12%**, and **24.38%**, respectively. This demonstrates the robust scalability of our hierarchical architecture to maintain both layout and identity control in intricate scenes. While some competitors show slightly higher scores on individual metrics

(a) Samples of basic part



(b) Samples of complex part

Figure 13: **Samples of the basic and complex part of IMIG-100K.**

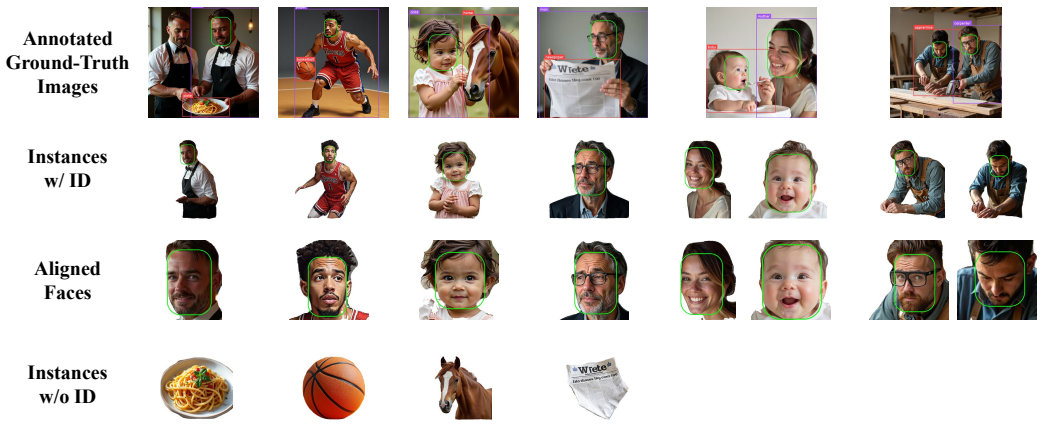


Figure 14: **Samples of the flexible composite part of IMIG-100K.** The faces are annotated in green round bounding boxes.

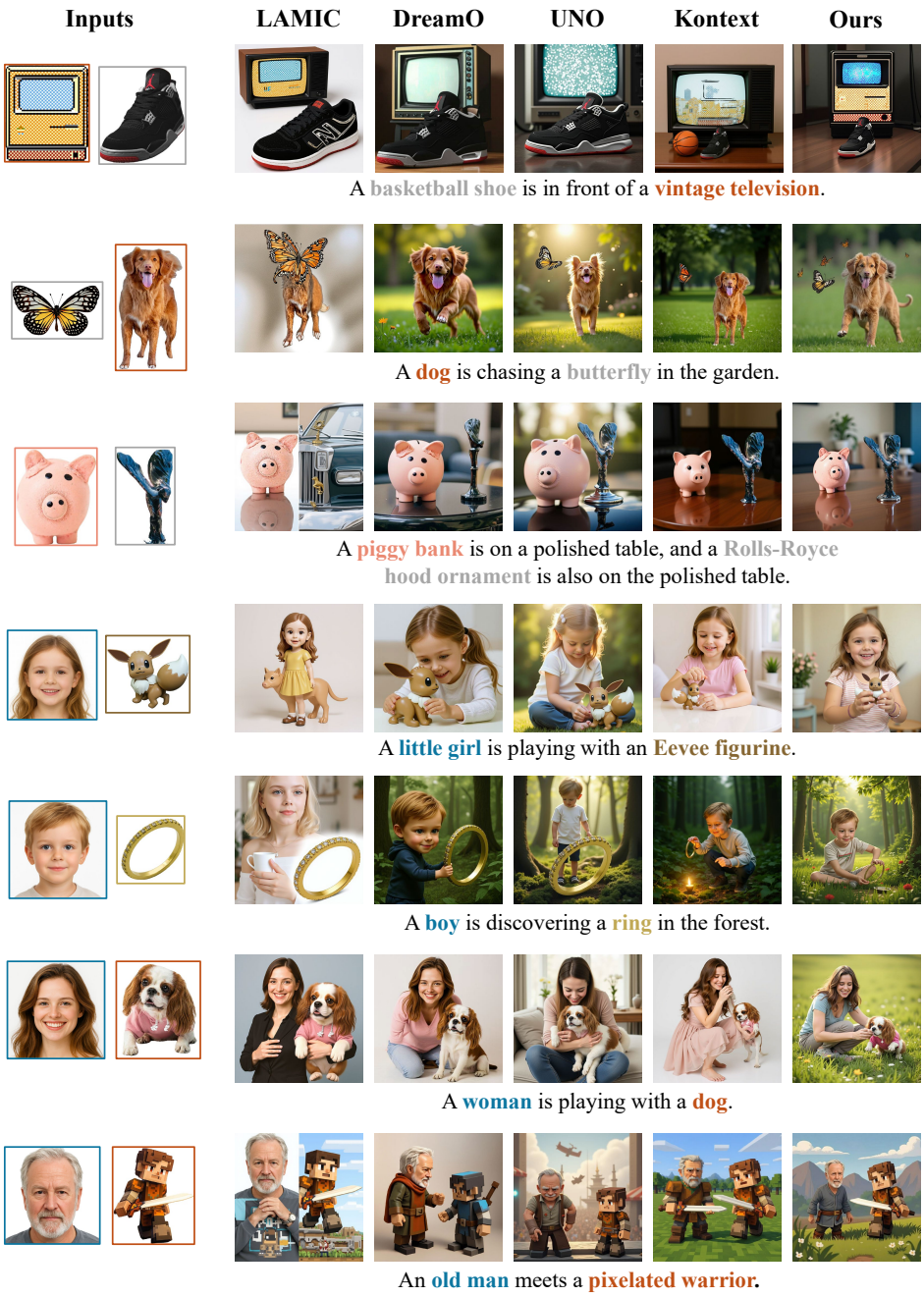


Figure 15: **Qualitative results across FLUX-based methods on LAMICBench++.**

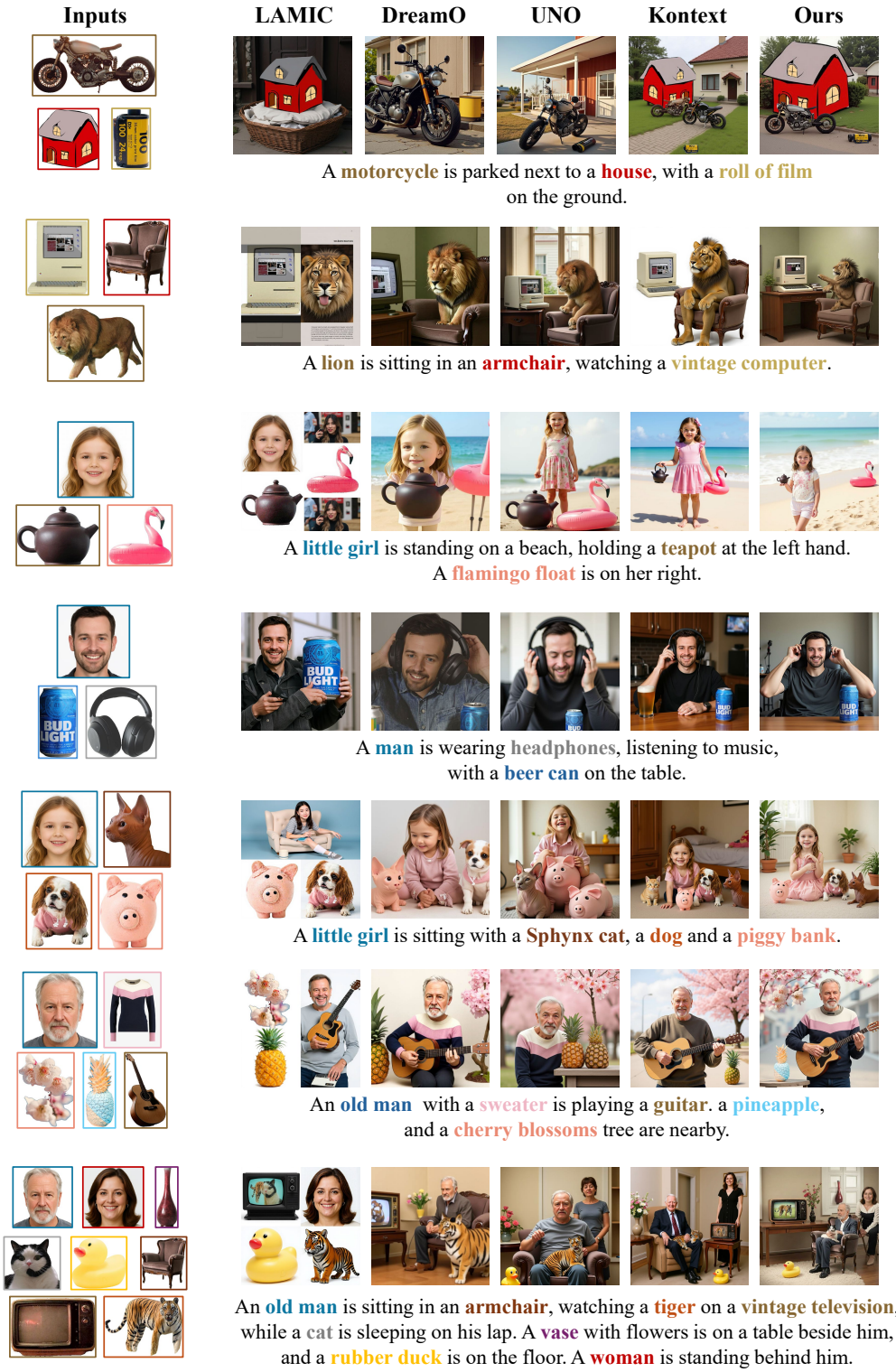


Figure 16: **Qualitative results across FLUX-based methods on LAMICBench++.**

Table 8: **Full quantitative result on COCO-MIG.** According to the count of generated instances, COCO-MIG is divided into five levels: L_2, L_3, L_4, L_5, and L_6. L_i means that the count of instances needed to generate in the image is i.

Method	Global Clip ↑	Local Clip ↑	Success Rate(%) ↑					
			Avg	L_2	L_3	L_4	L_5	L_6
LAMIC*	21.82	18.71	1.25	6.25	0.00	0.00	0.00	0.00
GLIGEN	25.21	20.90	4.25	16.88	4.38	0.00	0.00	0.00
MS-Diffusion*	25.50	20.77	4.50	13.75	5.62	2.50	0.62	0.00
CreatiLayout	**26.22**	20.70	19.12	46.25	30.63	11.88	4.38	2.50
3DIS	23.72	20.40	18.88	51.88	19.38	10.62	4.38	8.12
InstanceDiffusion	25.77	**21.91**	23.00	52.50	24.38	16.88	10.62	10.62
EliGen	24.92	20.58	26.00	50.00	**39.38**	22.50	10.00	8.12
MIGC	26.21	21.47	27.75	**53.75**	34.38	21.88	11.25	17.50
Ours*	25.86	21.87	**33.12**	52.50	37.50	**28.12**	**23.12**	**24.38**

Method	Instance Success Rate(%) ↑						mIoU ↑					
	Avg	L_2	L_3	L_4	L_5	L_6	Avg	L_2	L_3	L_4	L_5	L_6
LAMIC*	13.56	28.12	19.17	13.75	9.00	9.58	21.17	31.67	25.79	20.68	18.08	18.25
GLIGEN	29.56	41.88	31.67	27.19	27.38	27.81	27.44	37.35	29.17	25.31	26.42	25.56
MS-Diffusion*	28.22	37.81	33.12	28.12	25.75	24.69	34.69	41.15	36.38	34.57	32.36	33.70
CreatiLayout	54.69	67.19	63.33	56.09	50.25	48.96	48.96	56.32	55.38	49.42	46.22	45.28
3DIS	55.44	71.56	61.88	55.47	48.25	52.81	49.35	61.29	53.80	49.88	44.27	47.01
InstanceDiffusion	60.28	71.25	61.67	59.38	57.00	59.27	54.79	65.76	57.21	53.33	51.43	53.72
EliGen	64.12	69.69	**72.50**	66.56	61.62	58.54	59.23	64.61	66.10	61.59	56.74	54.50
MIGC	66.44	**74.06**	67.29	67.03	63.25	65.73	56.96	63.84	57.60	56.95	54.01	56.82
Ours*	**69.72**	70.94	69.58	**72.19**	**68.38**	**68.85**	**65.12**	**66.20**	**66.19**	**66.84**	**63.78**	**64.19**

like Global Clip (CreatiLayout (Wu et al., 2025c)) or Local Clip (InstanceDiffusion (Wang et al., 2024)), our approach achieves a superior overall balance. The consistently high mIoU scores across all complexity levels and our leading Instance Success Rate on the most challenging cases further validate our model's ability to master both precise instance placement and high-fidelity attribute preservation.

More qualitative results on COCO-MIG are shown in Fig. 17. Our method demonstrates a clear qualitative advantage over existing models on the COCO-MIG benchmark. In the first example, other methods fail to correctly generate the blue vase, with issues ranging from incorrect position to instance merging. Our model, in contrast, precisely renders the vase as intended. Similarly, for the "green potted plant", our approach correctly applies the "green" attribute to the pot, whereas competitors fail to do so. This highlights our superior ability to handle holistic subject identity. Furthermore, while some baselines correctly generate the requested objects in the third and fourth examples, their outputs often lack aesthetic harmony and visual coherence. Our method consistently produces images that are not only accurate but also visually pleasing and well-composed. These results underscore two key advantages of our framework: (1) Compared to other image-guided methods like MS-Diffusion (Wang et al., 2025), our approach offers significant superiority in layout control (2) Our dedicated identity preservation mechanism provides more robust and reliable subject fidelity than the attribute-based control of text-guided methods, particularly in intricate, multi-instance scenes.

D.3 More Qualitative Results on LayoutSAM-Eval

Additional qualitative results are presented in Fig. 18. Evidently, our method exhibits superior overall visual quality and realism compared to all existing approaches. While other methods (especially those reliant on text-guided layout-to-image generation) struggle with preserving fine-grained attributes—such as the specific text on the building in the first example and the exact color of the man's shorts in the second—our approach faithfully preserves the user's intended details to the greatest extent, excelling in both layout control and attribute binding.

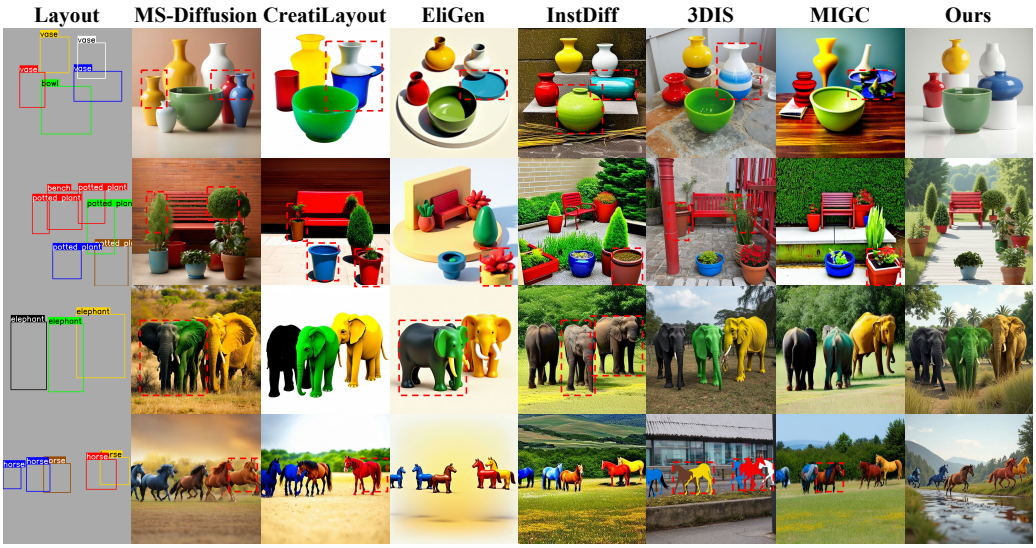


Figure 17: **More qualitative results on COCO-MIG.**

Caption: … with **large glass curtain walls** and **yellow** **text** that reads **"RYERSON UNIVERSITY"**… there are several trees with yellow leaves … surrounding environment includes other **buildings**, **shops**, and **vehicles**, as well as **pedestrians** on the street. **Phrases:** 1) Large **yellow** **text** on a glass building.. 2) A vibrant **yellow** **tree** stands out against a cityscape backdrop. 3) Modern **university building** with glass facade and distinct geometric patterns. 4) Lush **green** **tree** with **yellowing** **leaves**.

Caption: … where **a man in a blue uniform** is working in a park, pushing **a green wheelbarrow filled with branches**. **A deer** stands next to him, seemingly observing the man's actions … background is a lush green environment, with **trees** and **grassland**… **Phrases:** 1) **Man** in **navy blue** **shirt** and **khaki** **shorts** pushing wheelbarrow. 2) A **green** **wheelbarrow** with a **metal frame**, **black tires**, and visible wear. 3) A **deer** with **large antlers grazes on grass**. 4) A deer with **brown fur** and **white** **antlers**.

Figure 18: **More qualitative results on LayoutSAM-Eval.**

E LIMITATIONS AND FUTURE WORK

While our framework demonstrates state-of-the-art performance in multi-instance generation, it is not without limitations. A primary challenge stems from our model's strong emphasis on identity preservation. When inconsistencies exist between the provided reference images or between the images and the text prompt, our model tends to prioritize maintaining the identities of the reference subjects. This can sometimes lead to a lack of flexibility in adjusting attributes such as lighting, color, or pose, which may compromise the overall visual harmony and text-image consistency of the final output. This trade-off between identity fidelity and contextual flexibility represents an important area for future research. In the future, we plan to explore more dynamic attention mechanisms that can better balance these competing demands, allowing for more flexible style and attribute transfer. Drawing on zero-shot consistency priors (Zhang et al., 2025b), we also aim to incorporate broader generative knowledge to improve the overall harmony of multi-instance scenes. Furthermore, while ContextGen focuses on 2D image synthesis, extending identity-consistent control to 3D assets remains a promising direction (Xu et al., 2024; Shen et al., 2024).

F THE USE OF LARGE LANGUAGE MODELS

During the preparation of this manuscript, Large Language Models were used as a general-purpose writing assistant tool. Specifically, LLMs were employed to polish the language and refine the clarity of the text. The authors take full responsibility for the content of the paper.