SIGMA-GEN: STRUCTURE AND IDENTITY GUIDED MULTI-SUBJECT ASSEMBLY FOR IMAGE GENERATION

Anonymous authors
Paper under double-blind review

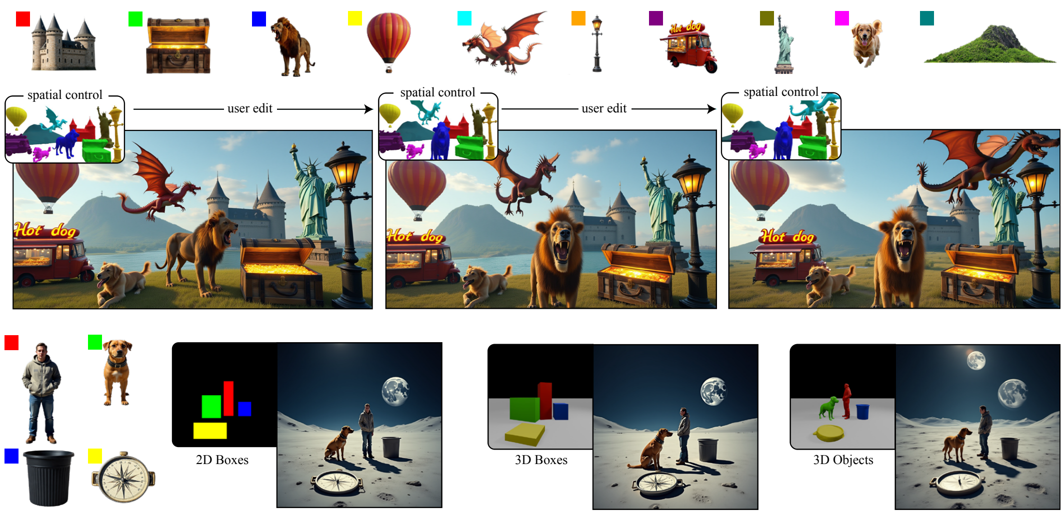


Figure 1: **SIGMA-GEN enhances controllability of text-to-image workflows by allowing users to prescribe both structure and subject identity.** In the top row, RGB images are used to describe subject identities. A 3D scene can be arranged by the user to describe the image structure; in these examples, meshes were automatically created using image-to-3D. The user can then assign identities to each subject (colors representing the assignments) and generate images while precisely editing the 3D scene. In the bottom part of the figure, we show that SIGMA-GEN can also be applied to simpler modes of structure guidance — 2D and 3D bounding boxes.

ABSTRACT

We present SIGMA-GEN, a unified framework for multi-identity preserving image generation. Unlike prior approaches, SIGMA-GEN is the first to enable single-pass multi-subject identity-preserved generation guided by both structural and spatial constraints. A key strength of our method is its ability to support user guidance at various levels of precision — from coarse 2D or 3D boxes to pixel-level segmentations and depth — with a single model. To enable this, we introduce SIGMA-SET27K, a novel synthetic dataset that provides identity, structure, and spatial information for over 100k unique subjects across 27k images. Through extensive evaluation we demonstrate that SIGMA-GEN achieves state-of-the-art performance in identity preservation, image generation quality, and speed.

1 INTRODUCTION

Recent advances in text-to-image generative models have enabled high-quality and diverse image synthesis from natural language prompts (OpenAI, 2023b; Rombach et al., 2022; Peebles & Xie, 2023; Lipman et al., 2022). However, these models still lack fine-grained control, which limits their adoption in real-world creative workflows. In particular, users have little ability to (i) control the identity of subjects and (ii) specify their arrangement within a scene.

We argue that both forms of control are essential. For identity, we introduce the use of single-view RGB images as descriptors, inspired by artistic workflows where exemplar images define visual elements for integration. For layout, we advocate for 3D object representations with rendered depth serving as a natural proxy for position, orientation, and occlusion relationships. To accommodate different levels of user expertise and control, we propose a single model that supports structural inputs at varying granularities: ranging from coarse 2D bounding boxes, to 2D masks, 3D bounding boxes, and per-pixel depth maps. This enables users to balance the ease of specification with the precision of control.

To the best of our knowledge, our work is the first to *jointly* support both structural guidance (capturing subject position, orientation, and occlusions) and identity guidance across multiple subjects within a single diffusion process. While prior works such as ControlNet (Zhang et al., 2023), Omini-Control v1/v2 (Tan et al., 2024), have demonstrated the use of multiple control modalities, none can reliably enforce multiple identities alongside structural layout guidance. Previous methods can be adapted to perform iteratively to address the issue of generating images with multiple subjects, but typically incur high runtime costs and suffer from compounding degradation in image quality. In contrast, our approach enables simultaneous generation of multiple identity-controlled subjects, arranged according to structural constraints, in a single forward pass.

The key to our approach is a large-scale synthetic data generation pipeline that automatically produces aligned RGB images, depth maps, masks, and identity descriptors. The pipeline leverages recent advances in image generation, grounding, and depth estimation. Using this pipeline, we construct a dataset containing **27k** images spanning over **100k** distinct identities, providing the supervisory signal required for training. We evaluate our models against competitive baselines and demonstrate significant improvements in both quality and controllability. When using per-pixel depth and masks, our method improves overall image fidelity by **31** points, identity preservation by **2** points, and achieves **4**× faster generation when synthesizing scenes with five or more distinct subjects. When analyzed under the coarser control modality of bounding boxes, our method improves overall image fidelity by **6** points and identity preservation by **11** points when synthesizing scenes with five or more distinct subjects. These results highlight the effectiveness of combining identity and structural guidance in a unified text-to-image generation framework.

2 Related Work

Structure control for image generation. Photorealistic image generation with diffusion models has paved the way for integrating conditions beyond text (Ho et al., 2020; Rombach et al., 2022; Podell et al., 2023; Peebles & Xie, 2023; Lipman et al., 2022). ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2024) introduce auxiliary conditioning pathways that enable structural guidance from inputs such as edges, depth, or segmentation. Other diverse of control such as bounding boxes (Li et al., 2023b; Zheng et al., 2023; Wang et al., 2024c; Cheng et al., 2024; Chen et al., 2023), and 3D priors (Bhat et al., 2024; Omran et al., 2025; Ma et al., 2023) have also been explored. However, these methods do not provide mechanisms for controlling subject identity.

Subject personalization. DreamBooth (Ruiz et al., 2023) and Textual Inversion (Gal et al., 2022) showed that diffusion models can learn subject identity by training a map of a few images of a given subject to a unique text token. Subsequent methods (Kumari et al., 2023; Gu et al., 2023; Liu et al., 2023a;b; Zhu et al., 2025; Jang et al., 2024) extend personalization to multiple subjects within a single image by incorporating positional information in text, adapting weights, or encoding spatial layout. However, all these methods require per-subject optimization to learn their appearance.

In-context generation. Another line of work trains diffusion models to incorporate subject images as conditioning (Ye et al., 2023; Wang et al., 2024a; Wei et al., 2023). With the advent of diffusion transformer (DiT) architectures, methods such as GPT1-Image (OpenAI, 2023a), Nano Banana (Google, 2025), and Flux Kontext (BFLabs et al., 2025) have been developed to support identity preservation. Recent open-source methods (Mou et al., 2025; Wu et al., 2025; Guo et al., 2025; Chen et al., 2025) extend identity preservation on diverse tasks controlled by text. However, these methods struggle to personalize across multiple subjects, as training datasets typically include only a small number of identities per image. Moreover, they lack support for structural control.

Subject insertion methods such as AnyDoor (Chen et al., 2024) and Insert Anything (Song et al., 2025) provide mask-guided control but are limited to inserting a single subject per generation. MS-Diffusion (Wang et al., 2024b) and other recent works (Tan et al., 2024; Tarrés et al., 2025; Li et al., 2023a; Xiao et al., 2025; Wang et al., 2025a) support different levels of spatial control, such as depth or bounding boxes. However, these methods are either limited to single-subject personalization or struggle with multi-subject personalization involving many identities. This can also be attributed in part to the scarcity of multi-subject training data. In contrast, SIGMA-GEN enables controllable multi-subject insertion, guided by identity images for each subject, within a single generation step (Figure 1). This leads to higher-quality and more coherent scene generation.

Subject personalization datasets. While several datasets support single-subject personalization– such as DreamBooth (Ruiz et al., 2023), CustomConcept (Kumari et al., 2023), AnyInsertion (Song et al., 2025), and Subjects200k (Tan et al., 2024)—only few focus on multiple subjects in an image. Virtual try-on datasets (Choi et al., 2021; Liu et al., 2016) contain 2–4 identities per image, including person identity and garments, but are limited to that domain. MultiWine (Tarrés et al., 2025) proposed a general-purpose dataset with up to two identities per image, obtained either from videos or through manual annotation. In contrast, we present the SIGMA-SET27K— a synthetically generated dataset with up to 10 subjects per image along with spatial controls and captions.

3 METHOD

Our objective is to generate an image I that includes subjects $s_1, s_2, \ldots s_n \in \mathbf{S}$ based on a prompt P, and controls C, such that all subject identities are preserved and are placed according to spatial controls C. To enable controllable generation with many subjects in one shot, we create a first-of-its-kind, high-quality dataset (§ 3.1) containing multiple subject identities in each image. Furthermore, we propose a lightweight representation (§ 3.3) that enables multi-subject generation effectively and efficiently. Finally, our dataset and proposed representation is used for finetuning models with varying granularity of structural control (§ 3.4).

3.1 SIGMA-SET27K DATASET

We illustrate our pipeline for generating SIGMA-SET27K in Figure 2. For each target image, our dataset provides per-subject data including an identity image, mask, depth and 2D/3D bounding box. We begin by prompting an LLM to produce an image-generation prompt describing multiple subjects against diverse backgrounds, along with subject and background captions. For each prompt, we generate the target image using an off-the-shelf text-to-image model. Next, we use a grounded segmentation tool to generate individual subject masks using the subject captions. We also employ a depth estimation model to predict the target image depth. Next, in a key step of our pipeline, we repose the subjects using Flux-Kontext (BFLabs et al., 2025) to obtain identity images with varying poses and lighting conditions. Finally, to enable coarser control than per-pixel depth, we fit 2D and 3D bounding boxes to the segmented subjects. Please refer to the Appendix for more details.

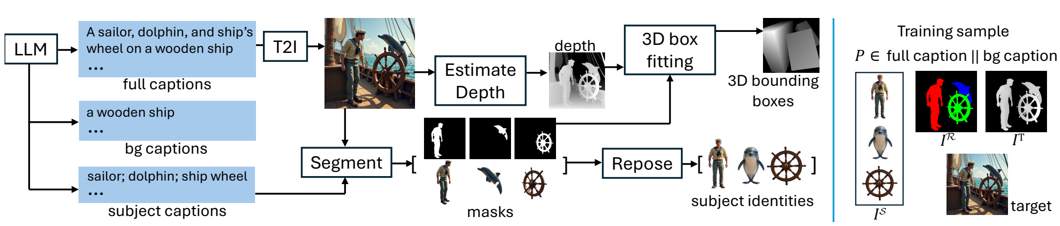


Figure 2: **Pipeline for generating SIGMA-SET27K.** Our fully automatic synthetic data generation pipeline involves creating compositional prompts with an LLM, generating images from these prompts, segmenting to obtain subject crops, reposing the crops to produce identity images, and estimating depth and 3D bounding boxes. We also show an example of a training sample for fine control scenario of using precise masks and depth. The routing mask is colored to RGB for visualization purpose, the pixel values for the subjects being 10, 20, 30 in practice for this example.

3.2 The SIGMA-Gen Architecture

SIGMA-Gen takes as input a prompt P, a multi-subject identity control image $I^{\mathcal{S}}$, and a spatial control image I^C to generate an image I. Both identity and spatial control images are encoded using the pre-existing VAE of a diffusion transformer model. Inspired by OminiControl (Tan et al., 2024), we adopt a unified attention mechanism in which the noisy image latents X and conditions $P, I^{\mathcal{S}}, I^C$ are concatenated along the token dimension as $[P, X, I^{\mathcal{S}}, I^C]$, enabling multi-modal attention from every modality to all modalities.

3.3 Representing Multi Subject Control

We decompose the spatial control C into two categories: routing control $\mathcal{R}$ and structure control T. The routing control specifies where each subject should be placed in the image, while the structure control can define additional overall control for the scene such as depth. To efficiently represent the spatial controls $C \in [\mathcal{R}, T]$, we embed them into a single spatial control image I^C with dimensions $H \times W \times 3$, matching the shape of the generated image I. We show an example of our control images in the right column of Figure 2, and in mask visualizations of Figure 3.

Let $\mathbf{S} = \{s_i\}_{i=1}^N$ denote the set of subjects and $\mathcal{R} = \{R_i\}_{i=1}^N$ denoting their desired spatial regions in the image domain. We define a mapping function $f : \mathbf{S} \longrightarrow \mathbb{M}$, which assigns each subject $s_i \in \mathbf{S}$ to a unique pixel intensity $m_i \in \mathbb{M}$. The control image $I^{\mathcal{R}}$ is constructed as $I^{\mathcal{R}}(x) = f(s_i)\mathbf{1}[x \in R_i]$. Thus, $I^{\mathcal{R}}$ with shape $H \times W$ encodes the subject layout, mapping each pixel to its corresponding intensity m_i, with 0 representing background. We discuss how we ensure each pixel maps to a single subject in the next section.

We now construct a subject identity condition image $I^{\mathcal{S}}$ (see training sample in right column of Figure 2). This image provides explicit identity information of each subject, allowing each pixel in the routing control to be unambiguously associated with its corresponding subject. Formally, let each subject s_i be represented by an image $I^{s_i} \in \mathbb{R}^{H' \times W' \times 3}$ of fixed spatial resolution $H' \times W'$. We then define $I^{\mathcal{S}}$ by concatenating all subject images $\{I^{s_i}\}_{i=1}^N$ along the height dimension which yields $I^{\mathcal{S}} \in \mathbb{R}^{(N \cdot H') \times W' \times 3}$. Thus the i-th $H' \times W'$ block of $I^{\mathcal{S}}$ corresponds directly to region $\mathcal{R}_i$, serving as a compact visual dictionary of subject identities and regions.

3.4 Multimodal Structure Control

We design a method capable of handling diverse control modalities—including pixel-level masks or depth maps, 2D bounding boxes, and 3D bounding boxes—within the same condition I^C, by concatenating the routing control $I^{\mathcal{R}}$ and the structure control I^T.

The routing control image $I^{\mathcal{R}}$ can be created from pixel-wise region information $\mathcal{R}$. However, the effectiveness of this control image relies on the precision of the masks. Accurate masks naturally handle occlusions and ensure that no two regions overlap. In contrast, coarser forms of control such as 2D or 3D bounding boxes, may produce ambiguous overlaps. In crowded scenes, this often results in partial or complete occlusion. To address this limitation while preserving an efficient representation in which all spatial conditions are encoded within a single image, we adopt a bidirectional compositing strategy. Specifically, we construct two routing control images $I^{\mathcal{R}_{\text{asc}}}$ and $I^{\mathcal{R}_{\text{dsc}}}$ by pasting the subject masks in ascending order of subject occurrence in $I^{\mathcal{S}}$ and descending order respectively. The mask values m_i associated with each subject s_i remain constant across both constructions; only the order of composition is varied. This bidirectional compositing increases the chances of a region being visible in either of the two routing control images. The structure control I^T refers to depth in our formulation and can be precise depths or coarser depths such as 3D bounding box depths. We refer the reader to Figure 4 for examples of coarser control images.

With the routing controls $I^{\mathcal{R}_{\text{asc}}}$ and $I^{\mathcal{R}_{\text{dsc}}}$ and structure control I^T we can create the spatial control image $I^C = \text{Concat}(I^{\mathcal{R}_{\text{asc}}}, I^{\mathcal{R}_{\text{dsc}}}, I^T; \text{channel}) \in \mathbb{R}^{H \times W \times 3}$. For the case where we have precise masks of each subject, both the routing control images are the same $I^{\mathcal{R}_{\text{asc}}} = I^{\mathcal{R}_{\text{dsc}}}$. This representation leads us to having only two condition images – a *subject identity control image* $I^{\mathcal{S}}$ and a *spatial control image* I^C which together can enable controllable multi-subject generation.

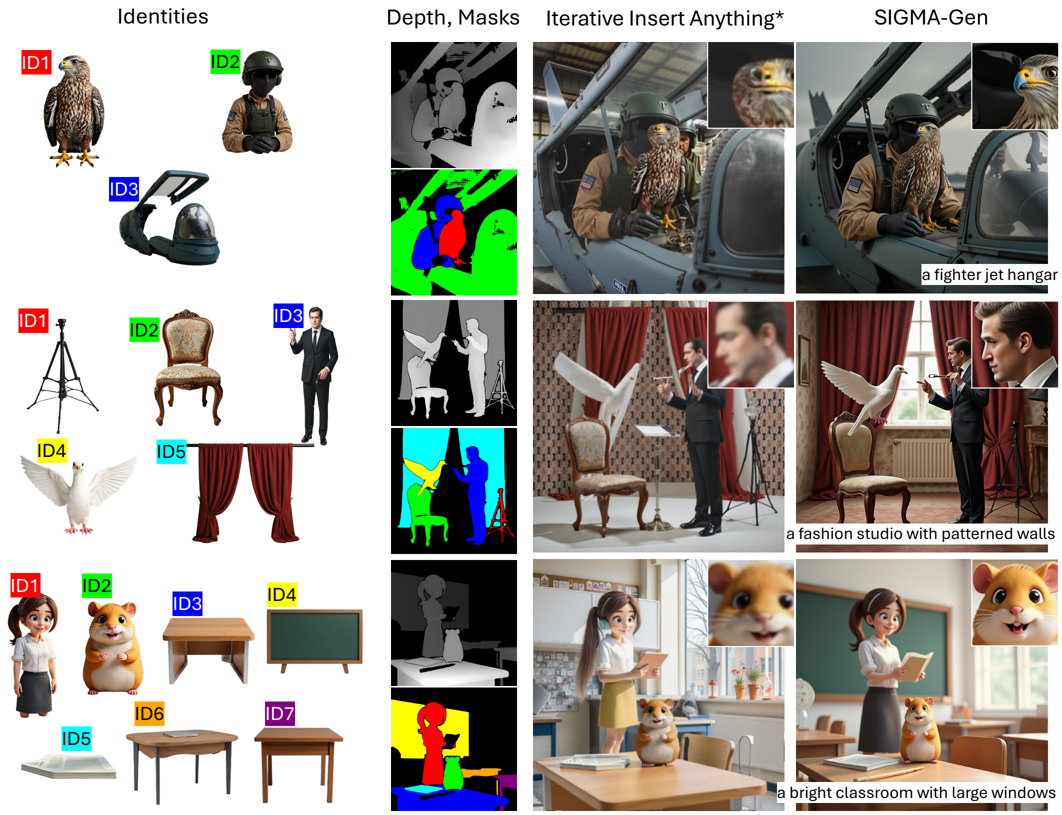


Figure 3: **Multi-subject generation with masks and depth.** SIGMA-GEN outperforms baselines both in terms of image quality (see zoomed crops at top-right) and subject identity preservation. For our case we prepend "Place these subjects in" to the prompts.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

We adopt Flux.1 Kontext [dev] as the base model for training our method. For spatial control, we use the same RoPE (Su et al., 2023) embedding as that of the noisy image. For identity control, we also use the same RoPE embedding but set the first dimension to ones (instead of zeros) to differentiate it from the noisy image, similar to Flux Kontext. Training is conducted on a single node with 8 A100 GPUs in three stages, using a LoRA rank and alpha of 128, the Prodigy optimizer (Mishchenko & Defazio, 2023), and a total batch size of 8. In the first stage, we train for 30k steps on a subset of data containing up to four subjects per image. The second stage trains for 20k steps on images with three or more subjects, followed by a third stage of 20k steps on images with more than four subjects.

To develop unified coarse-to-fine spatial conditioning, we randomly sample one of three structural inputs for each training example: (i) precise masks with depth, (ii) 3D bounding box masks with depth, or (iii) 2D bounding boxes. We further apply random dropping of one spatial condition channel with probability 0.1, as well as augmentations including random dilation of masks and bounding boxes and aspect ratio variation of bounding boxes by 1%, to improve robustness.

During training, we retain only the depths of the subjects, masking all other regions to zero. This mitigates the need to provide background depth at inference. For conditioning, we alternate between full prompts and background prompts with equal probability, constructing them as either *"Place these subjects in <bg prompt>"* or *"Place these subjects to compose: <full prompt>"*. For constructing the routing control image, we assign region intensity in steps of 10, with the region for the first subject marked with 10, the second subject with 20 and so on. We use non-gray colors while depicting masks in our figures for better visualization.

4.2 EVALUATION

For evaluation, we construct a dataset of 710 examples, including 200 single-subject personalized generation cases and 510 multi-subject cases. Dataset statistics are provided in the Appendix, and the data is generated as described in § 3.1.

We adopt DINO (Oquab et al., 2023) (DINO-I) and SigLIP (Zhai et al., 2023) (SigLIP-I) scores to evaluate subject identity preservation via image-to-image similarity. For multi-subject cases, we crop the generated image around the bounding box of each subject and compute similarity with the corresponding identity image, finally averaging over each crop and subsequently over every image. Cropping reduces the presence of other distracting elements in the image while computing similarity. We also evaluate text-to-image similarity using the SigLIP text-to-image score (SigLIP-T), which captures overall composition and background fidelity. We use SigLIP as it has been shown to outperform CLIP on fine-grained understanding tasks.

When precise depth is provided as control, we additionally compute the mean squared error (MSE) between the depth of the original and generated images, restricted to the subject regions. Unless otherwise specified, our evaluations use only subject depths as structure control. Finally, we assess perceptual quality with CLIP-IQA (Wang et al., 2023) and MUSIQ (Ke et al., 2021).

4.3 BASELINES

For single-subject personalization evaluation, we use OminiControl with its trained subject and depth LoRAs, as well as UniCombine. We also adapt a strong baseline, Flux Kontext, to incorporate depth guidance by attaching a depth ControlNet during inference. In addition, we design another strong baseline by combining the state-of-the-art mask-based insertion method, Insert Anything, with a depth ControlNet and a depth-to-image tool (BFLabs, 2024) to generate the initial image on which subjects are inserted. We refer to this variant as *Insert Anything** and use it for both single- and multi-subject personalization.

For the multi-subject setting, iterative insertion is the only feasible baseline, as no prior work enables multi-subject personalized generation with precise masks and depth. We also adopt MSDiffusion as a baseline for the coarser control task of generation with bounding boxes, for both single- and multi-subject cases. In our setup, bounding boxes are provided as filled mask images (see § 3.4), pasted in ascending and descending order of subject occurrence in the identity image. This design choice enables a unified control modality, whereas MSDiffusion instead constructs box embeddings directly from bounding box coordinates.

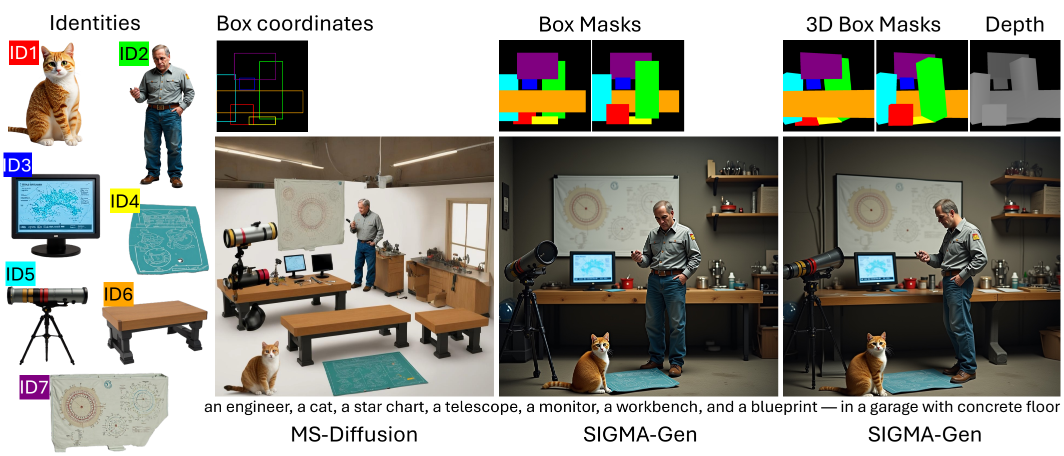


Figure 4: **Multi-subject generation with coarse controls.** Baseline fails to maintain position or identity, while SIGMA-GEN adheres to both 2D and 3D bounding-box coarse control. For our case we prepend “Place these subjects to compose: ” to the prompt.

	Method	Subject DINO-I	Subject SigLIP-I	Text SigLIP-T	Control	Depth MSE	Quality CLIP-IQA	Quality MUSIQ
Single Subject	UniCombine	75.89	82.72	15.52	Depth	103.1	**71.47**	70.81
	OminiControl	70.78	81.06	15.63	Depth	163.9	67.28	69.91
	Insert Anything*	74.69	81.44	**16.01**	Mask, Depth	216.2	65.97	67.35
	Flux Kontext*	79.96	83.22	15.99	Depth	112.4	66.30	64.93
	SIGMA-GEN	**81.04**	**83.99**	15.91	Mask, Depth	**70.2**	70.65	**70.87**
	MSDiffusion	74.83	82.65	2.41	Bbox	-	61.98	67.40
	SIGMA-GEN	**79.16**	**83.70**	**16.13**	Bbox	-	**70.82**	**70.75**
	SIGMA-GEN	79.98	84.00	16.03	Mask, 3D bbox	-	70.94	70.67
Multi Subject	Insert Anything*	72.72	75.58	17.66	Mask, Depth	203.4	44.41	48.86
	SIGMA-GEN	**74.54**	**77.82**	**17.73**	Mask, Depth	**26.35**	**72.64**	**73.21**
	MSDiffusion	63.28	69.06	11.20	Bbox	-	61.99	69.05
	SIGMA-GEN	**71.90**	**73.15**	**17.21**	Bbox	-	**68.83**	**70.96**
	SIGMA-GEN	73.48	75.27	18.19	Mask, 3D bbox	-	72.45	72.55

Table 1: **Quantitative comparison with baselines.** SIGMA-GEN achieves competitive or superior performance in single-subject controllable identity-preserving generation, and significantly outperforms baselines in the multi-subject setting.

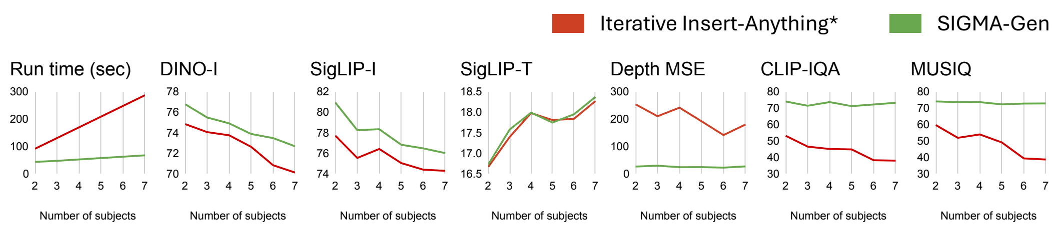


Figure 5: **Performance over increasing number of subjects.** Baseline *Iterative Insert Anything** run-time increases and quality decreases steeply compared to SIGMA-GEN. Our method also outperforms consistently in subject consistency and depth MSE.

5 RESULTS

We compare SIGMA-GEN with baselines both quantitatively and qualitatively. We further analyze different control types through ablations, showcasing emergent properties and applications.

5.1 COMPARISON WITH BASELINES

We present quantitative evaluation of our model with different types of controls in Table 1. Using a single trained model, we evaluate with (a) precise masks and depth, (b) 2D bounding boxes, and (c) 3D bounding box masks with depth. For this evaluation, SIGMA-GEN uses only subject depths, whereas other methods that accept depth are provided with full image depths. As shown in Table 3, SIGMA-GEN with subject depths performs comparably to the setting when full depth is provided. For the precise mask and depth case, we use background prompts for fair comparison to the baselines, while for the coarser controls we use full prompts as the bounding box baseline MSDiffusion expects full prompts too.

Our method consistently surpasses all baselines in subject consistency, as reflected by DINO-I and SigLIP-I scores for both single- and multi-subject personalization. For image-to-text similarity (SigLIP-T), our method outperforms all baselines except *Insert Anything** in the single-subject scenario. The stronger SigLIP-T performance of *Insert Anything** can be attributed to its use of a depth-to-image model with full depth to generate an initial image for subject insertion. Our ablation with full depth (Table 3) confirms that SigLIP-T benefits from access to full depth during generation.

Our method also achieves the lowest MSE between original and generated depths, demonstrating strong adherence to precise depth control while also supporting coarse depth guidance. In terms of

Controls	Subject DINO-I	Subject SigLIP-I	Text SigLIP-T	Depth MSE	Quality CLIP-IQA	Quality MUSIQ
Mask (BG)	74.17	77.52	17.62	40.10	71.26	72.27
Mask + depth (BG)	74.54	77.82	17.73	26.35	72.64	73.21
Mask + depth (FULL)	74.82	77.99	18.26	25.17	73.36	73.53

Table 2: **Ablation over increasing guidance.** (BG) represents text prompts describing only the background, whereas (FULL) describes the whole scene. Removing depth reduces performance while providing full prompts that include subject names improves performance.

Type of Depth Control	Subject DINO-I	Subject SigLIP-I	Text SigLIP-T	Depth MSE	Quality CLIP-IQA	Quality MUSIQ
Subject depths (tokens)	74.54	77.82	17.73	26.35	72.64	73.21
Full depth (tokens)	74.32	77.46	18.08	24.43	72.83	73.34
Full depth (ControlNet)	74.10	76.38	17.56	24.42	72.79	73.31

Table 3: **Ablation over depth control.** Providing depth beyond the subjects to our method leads to better quality, depth alignment, and text alignment. Providing the full depth using a ControlNet leads to loss in subject consistency, and text alignment.

image quality, measured by CLIP-IQA and MUSIQ, our approach is consistently competitive and often surpasses baselines across all control types. Notably, SIGMA-GEN performs strikingly well in the multi-subject setting, significantly outperforming all baselines across every evaluation metric.

For multi-subject personalization with precise masks and depth, we plot SIGMA-GEN's performance as the number of subjects increases (Figure 5), comparing against the strong baseline of *Iterative Insert Anything**. As expected, inference runtime grows much more steeply for the baseline due to repeated insertions. This iterative process also leads to a reduction in image quality, as reflected by lower CLIP-IQA and MUSIQ scores. In contrast, SIGMA-GEN maintains quality regardless of the number of subjects.

In terms of subject consistency, measured by DINO-I and SigLIP-I, SIGMA-GEN experiences some degradation as the number of subjects increases—largely due to dataset distribution effects (see Appendix)—yet still significantly outperforms the baseline. Text-to-image consistency scores remain similar between *Iterative Insert Anything** and SIGMA-GEN, with our method achieving higher performance on average. Finally, our approach preserves spatial consistency across subjects, as shown by stable MSE values, while the baseline's performance degrades with more subjects.

We show qualitative comparison for multi-subject personalization with precise masks and depth in Figure 3. We show examples with three, five and seven subject insertion. For the baseline, with three and five subjects, we see a loss of identity in the eagle and dove respectively. For seven subjects, along with loss of identity of the person we also observe that some subjects do not get inserted such as the chalkboard and the last two tables. In contrast, SIGMA-GEN successfully follows all spatial and subject controls.

We qualitatively compare the coarser tasks of 2D and 3D bounding box control against the baseline MSDiffusion, which also relies on bounding boxes. We show a case of seven subject insertion in Figure 4. MSDiffusion fails to maintain correct position or identity of the subjects. In contrast, SIGMA-GEN follows spatial and identity controls for both types of control granularities. We show more examples in the Appendix.

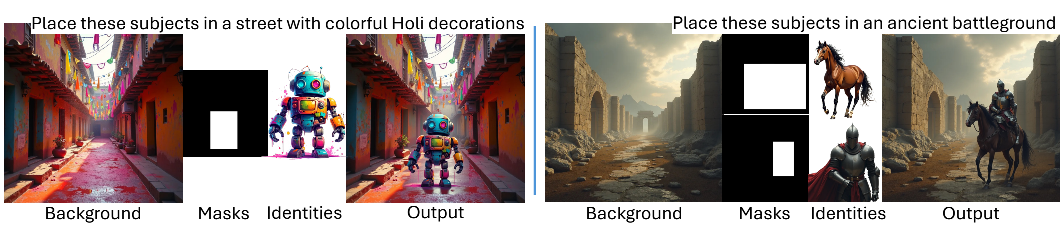


Figure 6: **Insertion using SIGMA-GEN.** We show extendability to one-shot single and multiple identity insertion given a background image based on bounding boxes.

5.2 ABLATIONS

In Table 2 we probe the performance of our method for the precise mask and depth scenario by 1) removing depth entirely while maintaining precise masks and 2) by providing full scene description as input prompt. Note than in Table 1 we used background prompts for the precise mask and depth case (also second row of Table 2) for fair comparison to baselines, but here we use full prompts

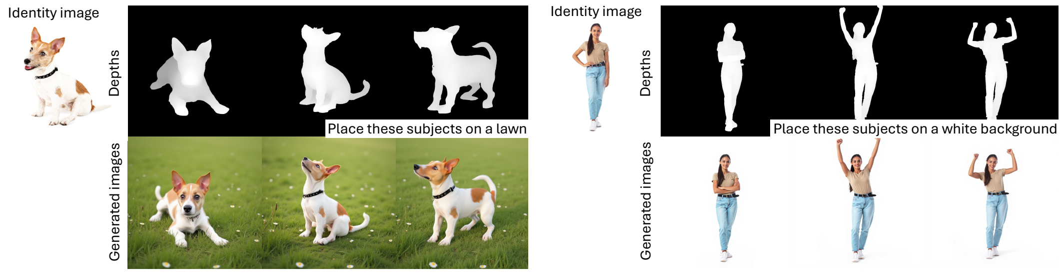


Figure 7: **Reposing subjects.** Through various poses supplied using depth, SIGMA-GEN can repose deformable subjects on different backgrounds.

for ablation in the last row. We observe that including the full prompt improves scores across all dimensions compared to using only the background prompt. We also observe that removing depth but providing precise masks (first row) reduces performance as less information is provided.

Secondly, in Table 3 we probe various methods of providing depth control to our trained model. We observe that providing the full depth improves performance in all aspects except for subject consistency where it decreases slightly. However, all scores consistently surpass the baselines. We perform an experiment where we omit depth by passing the structure control image I^T as all zeros, but pass depth externally through a pre-trained ControlNet. We observe degradation in subject consistency, text alignment and quality compared to providing the full depth via our structure control.

5.3 OTHER APPLICATIONS

Insertion. Although SIGMA-GEN has been trained for personalized image generation based on spatial/identity controls and prompt, we show that we can enable single and multi-subject one-shot insertion in Figure 6. For this task, we utilize blended diffusion (Avrahami et al., 2022) as a plug-and-play method to preserve the background. During inference, we follow the same strategy as we do for personalized generation, except we use the noised latents of the reference background image to preserve it . We show single and multi-subject insertion, where we observe that our method is able to harmonize the style of the subject to that of the background while maintaining correct shadows and lighting. It should be noted that the multi-subject insertion we show here is not iterative and is achieved in a single denoising loop.

Reposing. In Figure 7 we show control over a pose of an entity while maintaining its identity using depths of different poses. We show examples of deformable subjects whose identities across poses is more challenging to capture.

We show more emergent properties of SIGMA-GEN in the Appendix including the capability to handle free-form masks, style change, and multiple granularity levels of control in same generation.

6 CONCLUSION

We present SIGMA-GEN, a unified framework for controllable multi-subject, identity-preserving image generation. Our single model can follow spatial controls across fine-to-coarse granularities while preserving both the identities and arrangements of multiple subjects. To support this, we introduce a large-scale synthetic data generation pipeline that produces SIGMA-SET27K, a dataset with up to 10 subjects per image, which we use to train SIGMA-GEN. Through extensive evaluation, we demonstrate that SIGMA-GEN consistently outperforms baselines, with especially strong gains in scenarios involving five or more subjects. Finally, we showcase additional applications of SIGMA-GEN, like subject insertion and reposing.

7 REPRODUCIBILITY

We provide details of our method for synthetic data generation in § 3.1 and in the Appendix. We also state the implementation details of our method trained on said dataset in § 4.1. Our dataset, models and code will be open-sourced upon acceptance.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18208–18218, 2022.

BFLabs. Flux.1 depth [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-Depth-dev`, 2024. Accessed: 2025-09-22.

BFLabs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *arXiv preprint arXiv:2506.21416*, 2025.

Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. *arXiv preprint arXiv:2306.04607*, 2023.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6593–6602, 2024.

Jiaxin Cheng, Zixu Zhao, Tong He, Tianjun Xiao, Zheng Zhang, and Yicong Zhou. Rethinking the training and evaluation of rich-context layout-to-image generation. *Advances in Neural Information Processing Systems*, 37:62083–62107, 2024.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Google. Nano banana (gemini 2.5 flash image). Released August 2025, 2025. URL `https://www.google.com/gemini`. AI image generation and editing tool.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36:15890–15902, 2023.

Zinan Guo, Pengze Zhang, Yanze Wu, Chong Mou, Songtao Zhao, and Qian He. Musar: Exploring multi-subject customization from single-subject dataset via attention routing. *arXiv preprint arXiv:2505.02823*, 2025.

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

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *Advances in Neural Information Processing Systems*, 37: 100895–100937, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1931–1941, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023a.

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*, pp. 22511–22521, 2023b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023a.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023b.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

Wufei Ma, Qihao Liu, Jiahao Wang, Angtian Wang, Xiaoding Yuan, Yi Zhang, Zihao Xiao, Guofeng Zhang, Beijia Lu, Ruxiao Duan, et al. Generating images with 3d annotations using diffusion models. *arXiv preprint arXiv:2306.08103*, 2023.

Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023.

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, pp. 4296–4304, 2024.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025.

Mohamed Omran, Dimitris Kalatzis, Jens Petersen, Amirhossein Habibian, and Auke Wiggers. Controllable 3d placement of objects with scene-aware diffusion models. *arXiv preprint arXiv:2506.21446*, 2025.

OpenAI. Gpt-4 technical report. `https://arxiv.org/abs/2303.08774`, 2023a.

OpenAI. Dall·e 3. `https://openai.com/dall-e-3/`, 2023b. Accessed: 2025-09-22.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

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*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Wensong Song, Hong Jiang, Zongxing Yang, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in dit. *arXiv preprint arXiv:2504.15009*, 2025.

J Su, Y Lu, S Pan, A Murtadha, B Wen, and Y Liu Roformer. Enhanced transformer with rotary position embedding., 2021. *DOI: https://doi. org/10.1016/j. neucom*, 2023.

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

Gemma Canet Tarrés, Zhe Lin, Zhifei Zhang, He Zhang, Andrew Gilbert, John Collomosse, and Soo Ye Kim. Multitwine: Multi-object compositing with text and layout control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8094–8104, 2025.

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

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombine: Unified multi-conditional combination with diffusion transformer. *arXiv preprint arXiv:2503.09277*, 2025a.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025b.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024b.

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*, pp. 6232–6242, 2024c.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15943–15953, 2023.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

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*, pp. 3836–3847, 2023.

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 (CVPR)*, pp. 22490–22499, June 2023.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.

Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Xiu Li. Multibooth: Towards generating all your concepts in an image from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10923–10931, 2025.

A APPENDIX

A.1 ADDITIONAL DETAILS OF SIGMA-SET27K

We use Qwen-3-8B (Yang et al., 2025) for generating the full, background and subject captions. We prompt the LLM to generate image generation prompts with 3 to 10 subjects. Based on the individual subject captions, we use Grounded-Segment-Anything (Ren et al., 2024) to obtain segmentations of the subjects and masks. We apply multiple filtering criteria in this step. We remove any boxes that are less than 1% of the total image area and those that are greater than 40% of the image area. We also remove any duplicate or overlapping masks and only keep one of them. Finally, we only keep samples that have greater than 2 subjects per image. We use MoGe-2 (Wang et al., 2025b) for depth estimation. We estimated oriented bounded boxes using Open3D (Zhou et al., 2018) on depth for each subject. We repose each subject image using Flux.1-Kontext-dev to obtain identity images. Finally, we get 26435 images with a total of 105756 unique identities. We show the plot of our training data distribution in Fig. 8. For single and double subject data we process previously available datasets AnyInsert (Song et al., 2025) and MUSAR-Gen (Guo et al., 2025) to obtain spatial conditions and captions which we use only for the first stage of training, while using our dataset for the next two stages. For evaluation, since there exists no other dataset for multi-subject insertion, we generate the test set in a similar manner as described above. Our test set contains a total of 710 images and 2102 unique identities. We have 200 images containing one subject and we show the plot of the multi subject eval data in Fig. 8.

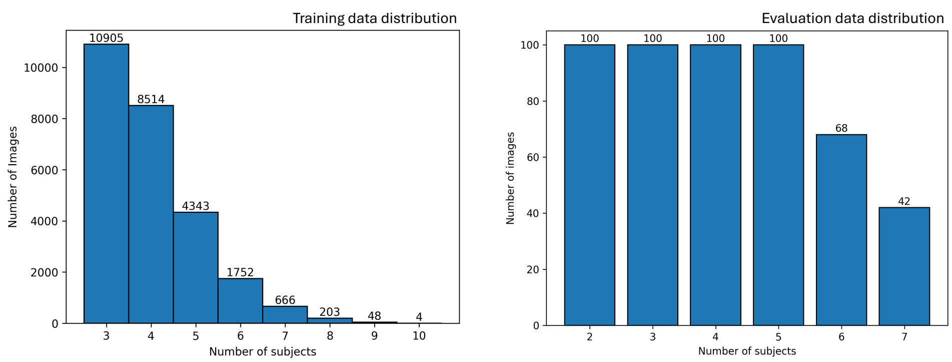


Figure 8: Train and test data distribution of number of images vs number of subjects they contain.

A.2 ADDITIONAL ANALYSIS

In Figure 9 we visualize an attention map of a region of the generated subject, marked in the depth map and the generated image. We find the attention between the noisy image and the identity image and visualize the attention mean over the chosen pixel areas for the full identity image. We observe that the attention map has higher activation near the pink tongue area of the bag that we marked on the depth/generated image. Identity was sampled from DreamBooth (Ruiz et al., 2023).

We also investigated how providing depth information for areas outside the the subject masks affects the results. Even though SIGMA-GEN has been trained only for subject depths whose masks are also provided, it has learned to decouple the depth from the masks. We show this in Figure 10 where we progressively add regions to the depth while maintaining the same mask and a single identity image. We observe that SIGMA-GEN follows the provided depth strictly in all cases.

A.3 ADAPTABILITY TO FREE-FORM MASKS DURING INFERENCE.

We probe the capability of our method to mask shapes unseen during training, such as circles or hand-drawn masks. We see in Fig. 11 that SIGMA-GEN can adapt to these forms of masks while maintaining identity and positions of subjects. Identities were sampled from DreamBooth.

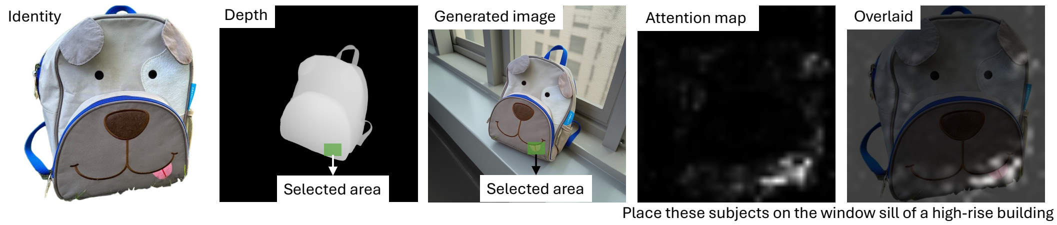


Figure 9: **Visualizing correspondence.** We visualize the attention map of a region marked in the noisy image space over the identity image.

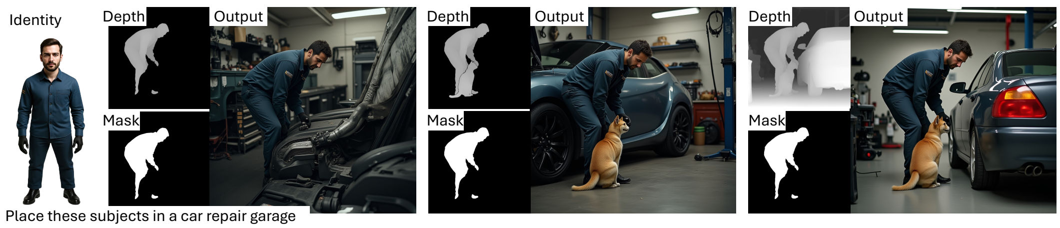


Figure 10: **Mask and depth are decoupled.** Despite being trained only with subject depths in the routing mask controls, SIGMA-GEN adapts effectively to increasing areas of depth control.

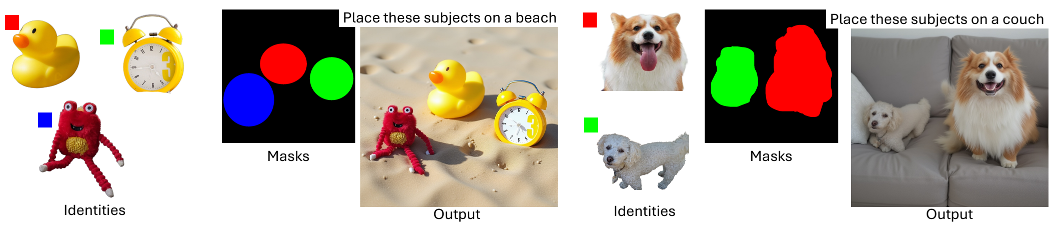


Figure 11: SIGMA-GEN can support circular or free-form masks during test time

A.4 STYLE CHANGE OF SUBJECT THROUGH PROMPT

Fig. 12 illustrates style changes, where prompts such such as "Claymation" and "pencil drawing" alter the style of the subjects while maintaining overall identity.

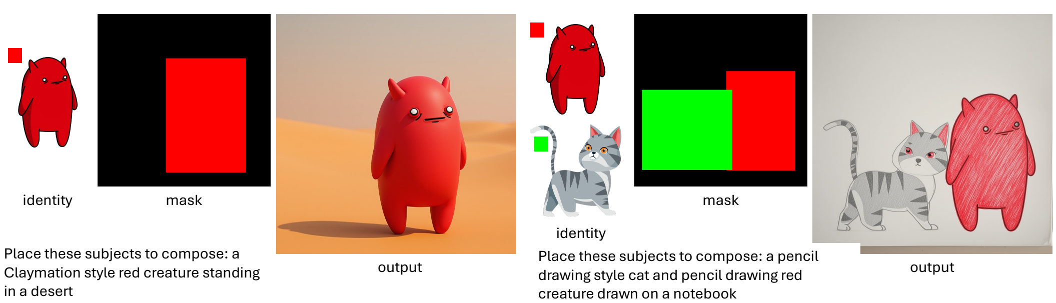


Figure 12: SIGMA-GEN can enable style change of subject(s) based on text.

A.5 COARSE TO FINE CONTROLS IN SAME GENERATION

In Fig. 13 we show that we can provide 2D boxes, 3D boxes, precise mask, and pixel-level depth at the same time for different subjects in the same image generation loop. This allows flexibility to choose the control level per subject. Identities were sampled from DreamBooth.

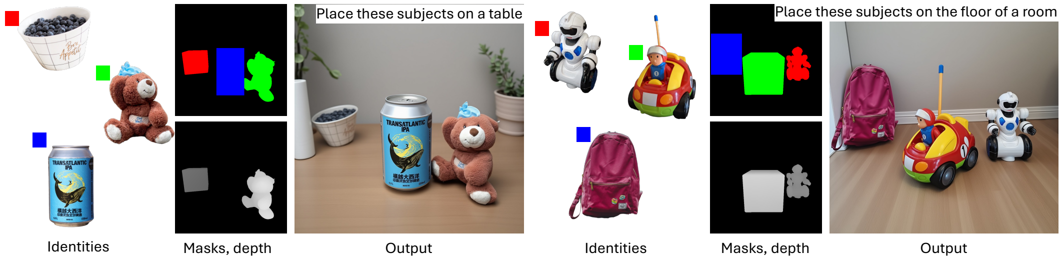


Figure 13: SIGMA-GEN can handle different types of conditions – pixel level mask, depth, 2D box, 3D box mask and depth – all at the same time.

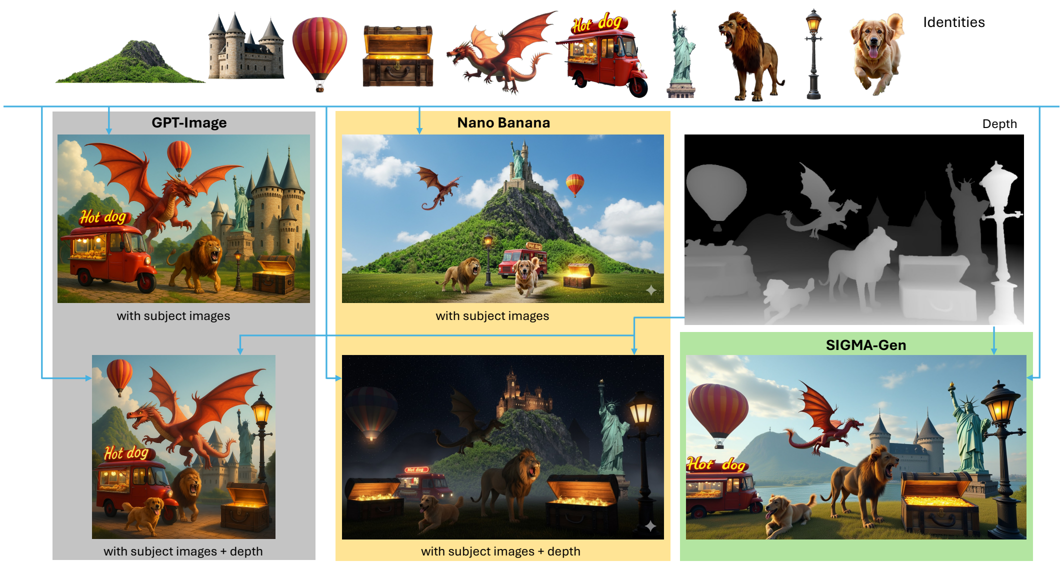


Figure 14: For 10 subject identity-preserved generation case, we test GPT-Image and Nano Banana.

A.6 10 SUBJECT IDENTITY-PRESERVED GENERATION

For the 10 subject generation scenario of Fig. 1, we use Nano Banana (Google, 2025) and GPT-Image (OpenAI, 2023a). GPT-Image generates an unnatural looking image, does not follow depth, and also leaves out a subject (dog) in the top image. Nano Banana generates a high quality image and can follow depth to some extent, however we see identity loss especially in the hot-dog cart, hot air balloon, and castle.

A.7 LLM USAGE

We use LLMs to find relevant prior work that could not be found with traditional search engines, to fix grammar and wording. We also used LLMs to brainstorm possible acronyms for the title. All outputs of the LLM were thoroughly verified by the authors before being added to the manuscript.

A.8 FAILURE MODES

In Figure 15, we show a couple of failure modes of our method. Firstly, for coarser controls if the overlap among regions is too high e.g. in the case of the cyan and yellow areas, the model may ignore one of the subjects. Secondly, if the viewpoint of the subject to be generated is significantly different from that of the identity image, the subject consistency may lower, e.g. observe the paws of the lion which seem to be facing the front. We also observe loss in human facial identity as the training data is not specifically designed for this task.

Figure 15: We show two failure cases of our method where subjects may be ignored in coarse control cases when the overlap is very high, and loss of subject consistency on high viewpoint change. On the example to the left, the significant overlap of the cyan box (scientist) and the yellow box (microscope) leads to the scientist not being generated. On the example to the right, even though the model generates a reasonable image with a viewpoint drastically different from the identity guidance, a closer look on the paws of the lion reveal they are actually pointing to the wrong directions (circled in red).

A.9 More comparison to baselines

We plot evaluation metrics over increasing number of subjects for box control scenario and compare with MSDiffusion in Fig. 16. We see that subject consistency degrades for MSDiffusion more steeply compared to ours as subjects increase as evidenced especially by DINO score. SIGMA-Gen also surpasses MSDiffusion across text-alignment and quality.

We show qualitative comparison with baselines for the single subject controllable identity-preserved generation case in Fig. 17 for precise mask and depth and in Fig. 18 for coarse controls – 2D and 3D box. SIGMA-Gen consistently preserves identity, generates high quality images and follows control unlike baselines.

We show more comparison with baseline for multi-subject identity-preserved generation for coarse controls in Fig. 19. Again we observe that the baseline loses out on identity, quality, and correct positioning, often ignoring some subjects. For SIGMA-Gen we see that the 3D box generations get more positioning information in terms of relative depths of subjects compared to the 2D box scenario and correctly positions all the subjects following the depth.

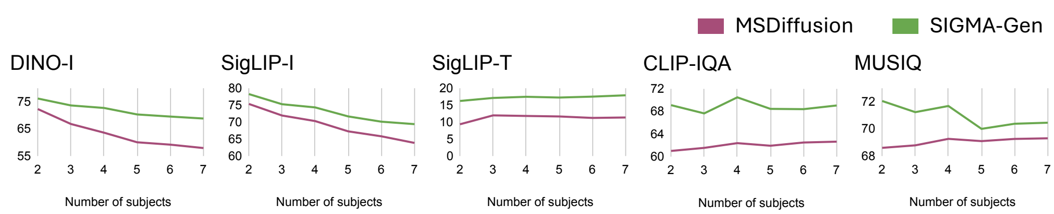


Figure 16: Comparison with baseline on coarse control multi-subject identity-preserved generation

Figure 17: Qualitative comparison with baselines for single-subject generation in the case of precise mask and depth control. We prepend “Place these subjects in” to the prompts for our method.

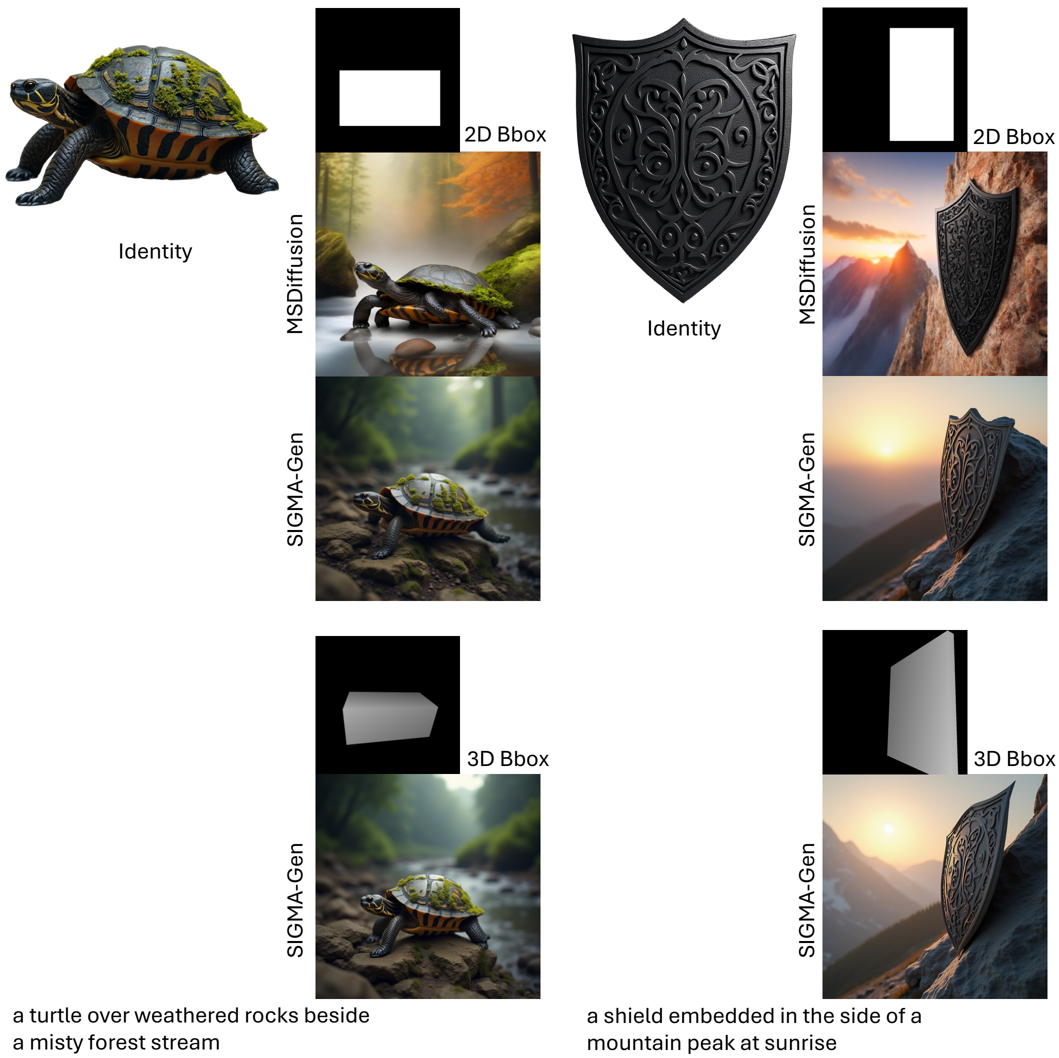


Figure 18: Qualitative comparison with baselines for single-subject generation in the case of coarse controls. We prepend "Place these subjects to compose: " to the prompts for our method.

Figure 19: More qualitative comparison results with baseline for multi-subject generation in the case of coarse controls. We prepend "Place these subjects to compose: " to the prompts for our method.

A.10 Post Reviews - Quantitative evaluation on DreamBooth

We conduct quantitative evaluation on the DreamBooth dataset Ruiz et al. (2023), where specifically we use the first image of every identity as the reference image and extract controls - depth and masks from all the other images which we treat as the targets. Similar to our data construction strategy described in Section A.1, we use MoGe-2 for depth estimation and Grounded-Segment-Anything for obtaining masks from both reference and target images. We also caption each target image using Qwen-2.5-VL. We use the same evaluation metrics as we did for our test datasets. In Table 4, we show that our method surpasses the previous baselines on the DreamBooth dataset as well.

	Method	Subject DINO-I	Subject SigLIP-I	Text SigLIP-T	Control	Depth MSE	Quality CLIP-IQA	Quality MUSIQ
Single Subject	Insert Anything*	77.95	81.60	16.22	Mask, Depth	201.7	67.18	67.77
	Flux Kontext*	82.29	82.91	16.26	Depth	99.8	67.05	64.99
	SIGMA-GEN	**82.84**	**84.06**	**16.27**	Mask, Depth	**58.4**	**71.43**	**70.32**
	MSDiffusion	76.44	82.63	2.07	Bbox	-	62.79	66.42
	SIGMA-GEN	**82.25**	**83.19**	**16.03**	Bbox	-	**71.50**	**70.19**
	SIGMA-GEN	82.30	83.88	16.17	Mask, 3D bbox	-	71.62	70.21

Table 4: **Evaluation on DreamBooth dataset for single subject generation.**

A.11 Post Reviews - Additional qualitative evaluation with real references

We would like to point to Figures 9, 11, 13 where we previously showed generations using DreamBooth identities. We also include additional qualitative results in Figure 20. For this figure, we sample identities from the Dreambooth and DeepFashion (Liu et al., 2016) datasets. Even though SIGMA-GEN has not been trained on datasets for human facial identity preservation, it handles person identities robustly. Performance on facial identity can be improved further with focused datasets in this domain. As stated in Section A.1, our training dataset also contains real domain single subject insertion - AnyInsert dataset samples which aids in robustness to real world samples.

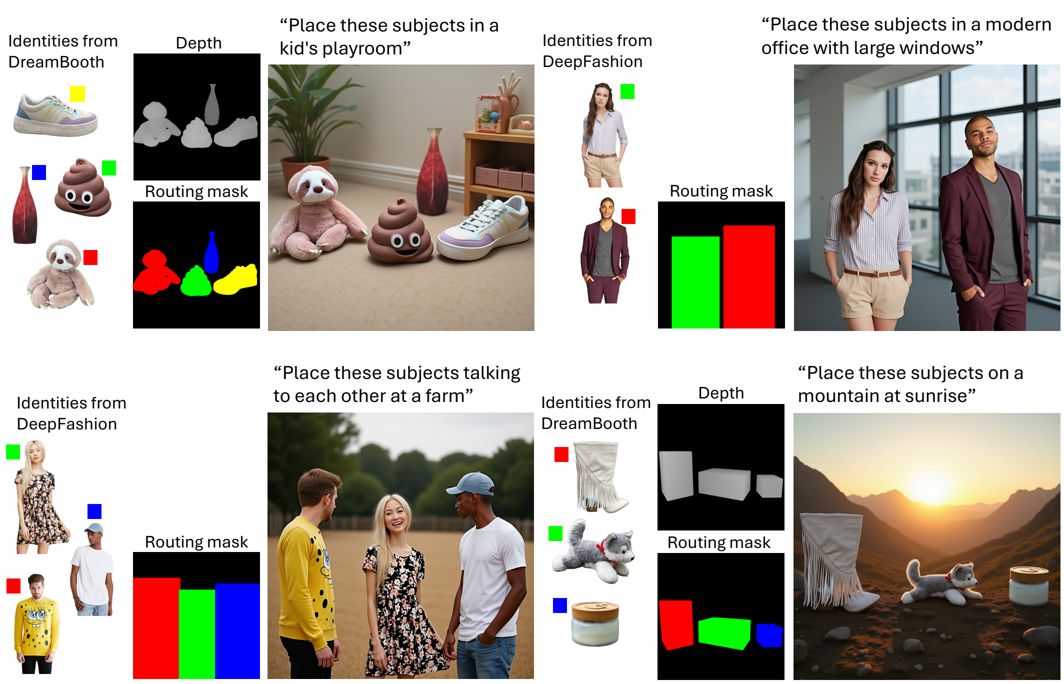


Figure 20: **Qualitative results with real reference images.** SIGMA-GEN robustly preserves identities of multiple real references across pose and viewpoint changes.

A.12 Post Reviews - Additional qualitative comparison with commercial models

We point to Figure 14 where we showed comparison to NanoBanana and ChatGPT for 10 subject controllable generation case. In Figure 21 we show that even with fewer number of subjects - 4 or 5 - both NanoBana and ChatGPT cannot adhere to the depth and also tend to lose identity (sloth toy in NanoBanana, vase in ChatGPT, red monster toy in NanoBana, IPA can in ChatGPT).

Figure 21: **Comparison to commercial models.** Unlike, NanoBanana and ChatGPT, SIGMA-Gen adheres to structure control and preserves identity of every subject.

A.13 Post Reviews - Comparison with Insert-Anything trained on our dataset

We fine-tuned Insert Anything starting with their given model on SIGMA-Set27K. We conduct this training by choosing one identity from our images at a time randomly to train for insertion using the same strategy as Insert Anything for 5k steps (as suggested in their paper). We show the results of this experiment in Table 5 (row 2) for multi-subject insertion. We see slight improvement in identity preservation, and more improvement in image quality. However, this model still significantly lags behind SIGMA-Gen mainly owing to the iterative strategy of insertion for multiple subjects.

Method	Subject		Text	Depth	Quality	
	DINO-I	SigLIP-I	SigLIP-T	MSE	CLIP-IQA	MUSIQ
Insert Anything*	72.72	75.58	17.66	203.4	44.41	48.86
Insert Anything^{TR}	72.84	75.65	17.64	187.3	49.93	52.06
SIGMA-Gen	74.54	77.82	17.73	26.4	72.64	73.21

Table 5: Effect of fine-tuning Insert Anything on SIGMA-Set27K

A.14 Post Reviews - Ablation over routing control

We explore different strategies for providing the routing control and show our ablation in Table 6. For the experiment with removing routing control (row 1), we omit the color assignment strategy and assign each box the same color intensity = 10. In row 2, we use the same order for pasting colored boxes for both channels (see Section 3.4). We observe that removing routing leads to a sharp decline in identity preservation. Removing biderection compositing also reduces identity preservation but less prominently, but image quality and text alignment remains similar. This can be associated to the model being able to generate most identities without bidirectional compositing but in erroneous locations and sizes.

Method	Subject		Text	Quality	
	DINO-I	SigLIP-I	SigLIP-T	CLIP-IQA	MUSIQ
without routing	64.07	64.69	17.11	68.11	69.43
without bidirectional compositing	69.67	71.92	17.19	68.82	70.42
proposed routing	71.90	73.15	17.21	68.83	70.96

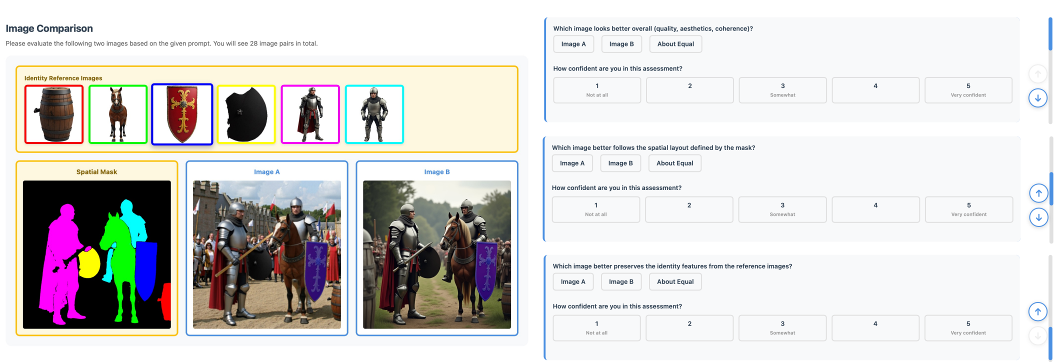
Table 6: Effect of routing strategy on multi-subject generation with 2D bounding boxes

A.15 Post Reviews - Ablation over effect of curriculum learning

We show in Table 7 that on training with the complete dataset together reduces performance across all aspects for multi-subject generation using 2D bounding boxes. Since we start training with Flux Kontext which can handle a single reference image, we find it beneficial to gradually increase number of references in terms of reducing routing confusion and ensuring all subjects get included.

Method	Subject		Text	Quality	
	DINO-I	SigLIP-I	SigLIP-T	CLIP-IQA	MUSIQ
without curriculum learning	65.36	66.39	17.04	67.94	68.47
with curriculum learning	71.90	73.15	17.21	68.83	70.96

Table 7: Effect of curriculum learning on multi-subject generation with 2D bounding boxes



a) visualization b) questions

Figure 22: **Setup of human study.** We illustrate our designed human study setup. a) We show identities, routing mask and generated images of Insert-Anything* and SIGMA-Gen. On hovering over an identity (shield selected in this case) the corresponding region in the generated images is highlighted for the users to compare. b) We show the questions we ask the users to evaluate the methods.

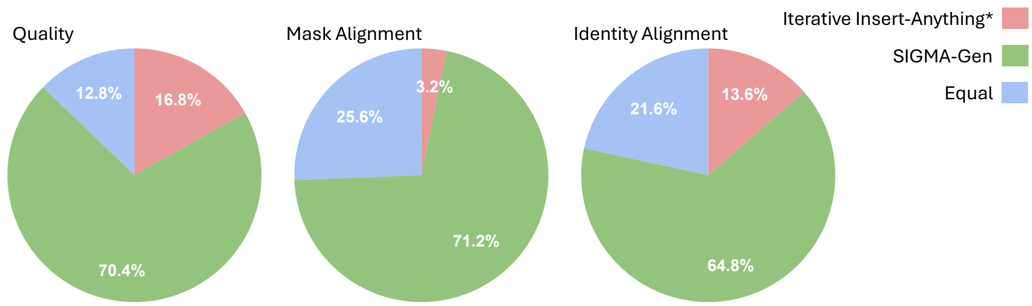


Figure 23: **Results of user study.** We show that users prefer SIGMA-GEN over the baseline more frequently in terms of image quality, alignment to mask control, and alignment to identity control.

A.16 POST REVIEWS - HUMAN STUDY

We show the setup of our designed human study in Figure 22. We designed this evaluation for using precise mask and depth with SIGMA-GEN and the baseline Insert-Anything* for multi-subject identity preserved generation. Our user study contains 30 samples of multi-subject generation - 5 each for two, three, four, five, six, and seven subject cases. Up till December 2nd, we have obtained a total of 125 responses across 19 participants. We show the results in Figure 23. We observe that users preferred SIGMA-GEN in all three areas considered - quality, mask alignment and identity preservation.