CanvasComposer: Personalized Group Photo Generation via a Multi-Reference Canvas

GORDON GUOCHENG QIAN*, Snap Inc., USA
RUIHANG ZHANG*, Snap Inc., USA and University of Toronto, Canada
TSAI-SHIEN CHEN, Snap Inc., USA and University of California, Merced, USA
YUSUF DALVA, Snap Inc., USA and Virginia Tech, USA
ANUJRAAJ GOYAL, Snap Inc., USA
WILLI MENAPACE, Snap Inc., USA
IVAN SKOROKHODOV, Snap Inc., USA
MENG DONG, Snap Inc., USA
ARPIT SAHNI, Snap Inc., USA
DANIIL OSTASHEV, Snap Inc., USA
JU HU, Snap Inc., USA
MUKESH SINGHAL, University of California, Merced, USA
SERGEY TULYAKOV, Snap Inc., USA
KUAN-CHIEH JACKSON WANG, Snap Inc., USA

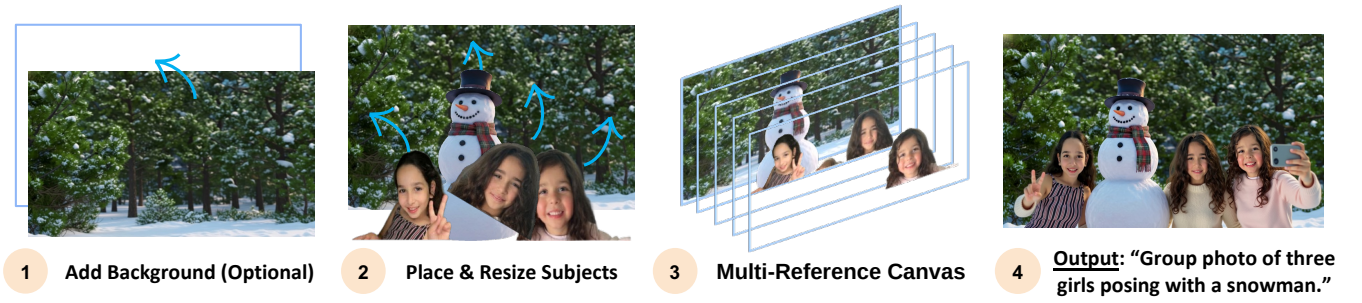

Fig. 1. **CanvasComposer** introduces an **interactive personalization** paradigm for multi-human text-to-image generation. It gives users a Photoshop-like composition experience: reference subjects are placed on a **multi-reference canvas**, with each identity individually adjustable through direct canvas manipulation. CanvasComposer then harmonizes the multiple references into a coherent, high-fidelity group image that follows the prompt.

Existing personalized image generators still struggle to preserve multiple reference identities in natural and coherent multi-human generations. To address these limitations, we present CanvasComposer, an interactive framework for personalized group photo generation. Inspired by professional image-editing software, CanvasComposer allows users to place reference subjects on a shared canvas, where each subject keeps its own RGBA cutout of the input. This multi-reference canvas preserves reference content under overlap while providing an intuitive interface for organizing multiple identities; the subjects remain separate elements on the input canvas, and the model outputs a single personalized and harmonized image. To keep this representation efficient, transparent latent pruning retains only tokens from each subject's non-transparent region, and cross-reference training mitigates copy-paste artifacts by learning to harmonize references sampled from different images. Extensive experiments demonstrate that CanvasComposer achieves coherent generation and strong identity preservation compared to state-of-the-art methods in multi-human personalized image generation.

CCS Concepts: • **Information systems** → **Multimedia content creation**; • **Computing methodologies** → **Image processing**.

*Equal contribution. ule[-10pt]0pt10pt

Additional Key Words and Phrases: Personalized Text-to-Image, Multi-Human Composition, Human Image Generation, Generative Model, Diffusion Model

1 INTRODUCTION

The advent of modern text-to-image (T2I) diffusion models [Rombach et al. 2022] has marked a pivotal moment in digital content creation, enabling the synthesis of complex, high-fidelity images from simple textual descriptions. This breakthrough has spurred a wave of research into personalization, enabling content creation with specified human identities in unseen contexts. Textual Inversion [Gal et al. 2023], DreamBooth [Ruiz et al. 2023], and IP-Adapter [Ye et al. 2023] have made significant progress in this direction.

These successful prior works [Guo et al. 2024; Ruiz et al. 2023; Ye et al. 2023] primarily target single-identity personalization. However, personalization is often social: users may want to generate images that include themselves together with friends, family, or other known people, making multi-human personalization a fundamental use case. A successful model must preserve each reference

Fig. 2. **Motivation for the multi-reference canvas in multi-human personalization.** State-of-the-art multi-human personalization methods such as UniPortrait [He et al. 2025] and OmniGen2 [Wu et al. 2025c] often struggle in challenging four-person scenes, where reference identities may be swapped, duplicated, or omitted, and conditioning cost grows with the number of subjects. In contrast, CanvasComposer introduces a multi-reference canvas where users place and adjust each reference subject's relative location through direct canvas manipulation, leading to stronger identity preservation, generation quality, and user experience.

identity, bind each reference to the correct generated person, and render all subjects coherently in a shared scene.

Existing multi-human personalization methods [He et al. 2025; Qian et al. 2025a; Wu et al. 2025c,a; Zhang et al. 2025b] typically encode each reference subject into visual conditioning tokens using CLIP- [Radford et al. 2021], face- [Deng et al. 2022; Qian et al. 2025b], or VAE-based encoders [Kingma and Welling 2013], concatenate the tokens from all subjects, and inject them into a pretrained diffusion model through decoupled cross-attention [Ye et al. 2023] or in-context conditioning [Wu et al. 2025c; Zhang et al. 2025b]. This token-concatenation paradigm faces two key challenges. First, subject association becomes ambiguous as the number of people increases: the model must infer which reference corresponds to which generated person, often causing identity swaps, identity mixing, duplicated subjects, or omissions. Second, the conditioning sequence grows linearly with the number of personalized subjects, increasing computation and memory. As a result, state-of-the-art multi-person personalization methods struggle in four-person settings, as demonstrated in Fig. 2. These limitations motivate a new paradigm for *interactive personalization with efficient multi-identity handling*.

To meet this demand, we introduce **interactive personalization**, where users directly compose an input canvas for reference-based group photo generation. Inspired by professional editing tools (e.g., Photoshop), our framework empowers users to act as design directors: they can intuitively compose a scene by placing multiple reference subjects on a multi-reference canvas and adjusting their positions and sizes, as shown in Fig. 1 (1–2). The canvas keeps the subjects separated, allowing users to adjust their relative locations through simple layout manipulation while combining an optional background and multiple identity references in a single multi-reference input. The model renders these references into a coherent, high-fidelity, personalized group photo. Throughout this paper, each subject or background is placed on the canvas as its own reference image with no fixed ordering among them: these references may overlap freely and serve only to specify the personalized subjects for multi-human generation, and the model outputs a single flattened image.

To render such a template, we propose ***CanvasComposer***, a generative framework built on the pretrained FLUX Kontext [Black Forest Labs 2025] diffusion transformer and designed specifically for interactive multi-human T2I generation. The core of CanvasComposer is the **multi-reference canvas**, an input representation where each subject is represented by its own RGBA cutout shown in Fig. 1 (3). Rather than compositing references into a single flattened input image, the multi-reference canvas keeps subjects separated in the input while preserving the user-edited canvas layout. The multi-reference representation offers two critical advantages. First, it naturally handles input-side occlusion by keeping each subject as a separate RGBA cutout, preserving subject content even when subjects overlap in the input canvas. In contrast, a simple collage-based approach inevitably introduces occlusions when multiple subjects overlap within the input canvas. Such occlusions result in partial identity information loss, and when adjustments are required, such as aligning subjects to match the text prompt or enabling natural interactions, the model inevitably hallucinates the missing regions, leading to reduced identity fidelity (Fig. 9). Second, we introduce a **transparent latent pruning** strategy demonstrated in Fig. 4, which extracts and concatenates only valid (non-transparent) tokens from each subject. By discarding empty transparent regions, this design makes the conditioning length depend on non-transparent reference content rather than the number of subjects placed on the canvas, eliminating the linear per-subject token growth of prior concatenation-based methods and improving efficiency for multi-human generation.

To train CanvasComposer, the conventional training pipeline constructs such a canvas by cropping human segments directly from the target image to place on the canvas (Fig. 3 right). However, this approach often results in undesirable copy-paste artifacts due to pixel-level correspondence between the input and target. To address this, we introduce **subject-wise cross-reference training** (Fig. 3 left). We curate a multi-image-per-scene dataset in which each identity group contains multiple images. During training, an image is randomly chosen as the target, while the corresponding humans and background in the multi-reference canvas are sampled from other source images within the same image set. Each sampled human is fitted to the target bounding boxes, with additional geometric and lighting augmentations to introduce intentional mismatches between inputs and targets. The subject-wise cross-reference training strategy encourages the model to disentangle redundant factors such as pose and illumination across subjects, thereby harmonizing multiple inputs from diverse contexts into a single coherent image during inference.

Our main **contributions** are as follows:

- We propose an **interactive personalization paradigm** that allows users to place multi-human references in full-body, portrait, or cropped-head forms on a multi-reference canvas.
- We introduce the **multi-reference canvas**, a simple and effective input representation that handles occlusions in multi-human personalization and improves efficiency through our **transparent latent pruning** strategy.
- We present **the subject-wise cross-reference training strategy**, which effectively disentangles redundant visual information (e.g., poses) and mitigates copy-paste artifacts.

- We develop **CanvasComposer**, a practical framework for interactive multi-human personalized generation. Extensive experiments demonstrate that CanvasComposer achieves state-of-the-art identity preservation and generation quality across multi-human personalization benchmarks.

2 RELATED WORK

Personalized Generation. Personalization methods have evolved from costly per-concept tuning [Gal et al. 2023; Kumari et al. 2023; Nitzan et al. 2022; Ruiz et al. 2023] to recent adapter-based solutions [Gal et al. 2024; Goyal et al. 2025; Guo et al. 2024; Han et al. 2024; Li et al. 2024; Patashnik et al. 2025; Qian et al. 2025b; Wang et al. 2024a; Ye et al. 2023], which enable efficient personalization while keeping the base diffusion model frozen. For multi-subject personalization, optimization-based approaches require additional concept disentanglement [Avrahami et al. 2023; Garibi et al. 2025; Kumari et al. 2023] or training multiple LoRAs [Dalva et al. 2025; Kong et al. 2024; Po et al. 2024]. Optimization-free approaches [Chen et al. 2025; Kim et al. 2024; Wang et al. 2024b; Xiao et al. 2025b; Zhou et al. 2024], including recent works such as UniPortrait [He et al. 2025], ID-Patch [Zhang et al. 2025b], ComposeMe [Qian et al. 2025a], and MS-Diffusion [Wang et al. 2025], rely on lightweight adapters but become increasingly expensive as the number of subject tokens grows. Recent in-context models [Comanici et al. 2025; Mi et al. 2025; Mou et al. 2025; Wu et al. 2025b,a; Xiao et al. 2025a] support multiple concept combinations, but they provide limited interactivity and their cost still grows with the number of subjects. Our CanvasComposer instead advances this line of work by introducing a multi-reference canvas that supports interactive subject placement, occlusion-aware reference preservation, and efficient multi-human personalization.

Structural and Compositional Conditioning. A broad range of conditioning mechanisms has been proposed to guide T2I models with additional visual or structural inputs. ControlNet [Zhang et al. 2023] and T2I-Adapter [Mou et al. 2024] are pioneering works that inject structural cues through auxiliary components, allowing the diffusion model to incorporate signals such as pose or depth. GLIGEN [Li et al. 2023b], LayoutDiffusion [Zheng et al. 2023], and their follow-ups [Chen et al. 2024; Cheng et al. 2024; Song et al. 2023; Xiong et al. 2024; Zhang et al. 2025a] fine-tune diffusion models to interpret bounding boxes or segmentation masks for structured generation. Training-free approaches such as NoiseCollage [Shirakawa and Uchida 2024], GrounDiT [Lee et al. 2024], and Bounded Attention [Dahary et al. 2024] manipulate noise or attention maps at inference. In terms of layered image generation, CollageDiffusion [Sarukkai et al. 2024] and LayerDiffusion [Li et al. 2023a] are leading works, which require per-layer optimization through textual inversion [Gal et al. 2023] to keep layers separated. SceneDiffusion [Ren et al. 2024] extends LayerDiffusion to remove the optimization stage but still requires per-layer generation for image editing. Latent transparency [Zhang and Agrawala 2024] and RGBA-instance generation [Fontanella et al. 2024] instead produce transparent layers on the *output* side, and LaRender [Zhan and Liu 2025] controls output occlusion training-free by volume-rendering prompt-described objects in latent space; none of these condition on reference images for identity personalization. In contrast to these box-, attention-, or RGBA-*output*-based mechanisms, CanvasComposer takes a separate RGBA image per subject as multi-reference *input* conditioning directly, preserving reference identities even when these input images overlap; concurrent layered-generation efforts are complementary to this input-side design. LazyDiffusion [Nitzan et al. 2024] is also related in spirit because it reduces diffusion-transformer computation for localized editing. However, it prunes target denoising tokens inside a user-specified mask, whereas CanvasComposer prunes input conditioning tokens from transparent regions across multiple reference images. The most relevant work to ours is ID-Patch [Zhang et al. 2025b], a feedforward personalized T2I model trained on collaged face inputs, where each patch consists of projected face features optionally overlaid with pose. However, ID-Patch is restricted to face-only generation and suffers from occlusion issues as well as copy-paste artifacts. In contrast, CanvasComposer organizes full-body, portrait, or cropped-head references as separate input images, preserves overlapping inputs, and effectively mitigates copy-paste artifacts via the proposed subject-wise cross-reference training.

3 CANVASCOMPOSER

3.1 Multi-Reference Canvas

CanvasComposer is a multi-human text-to-image generation framework that offers an interactive personalization experience, enabling users to specify the appearance of multiple subjects (humans and an optional background) through separate visual references. Concretely, the framework conditions a pretrained diffusion model on two inputs: (1) a text prompt that specifies global image content and high-level semantics and (2) a **multi-reference canvas** that keeps each subject as a separate reference element. This design supports intuitive canvas layout manipulation while keeping identity-specific guidance separated in a globally coherent image.

Illustrated in the second column of Fig. 3, the **multi-reference canvas** is represented by a set of RGBA images $L = \{l_1, \cdots, l_N\}$, one per subject or background element, where N is the number of such elements. Each RGBA image l_i represents one element, either a human segment or a background image. The RGB channels provide visual reference for the element, while the alpha channel defines valid regions. This mask is used to identify valid tokens, while invalid tokens are pruned out for efficiency, as detailed in Sec. 3.2. When an optional background is provided, regions covered by foreground subjects are removed from the background during both training and inference, letting the diffusion model harmonize the final image according to the text prompt.

Each subject or background is placed separately on the canvas with no fixed ordering among them; a per-element index only keeps overlapping references separable (Sec. 3.2) and carries no depth semantics. The alpha channel is used to prune input tokens for efficiency, and the output is a single flattened image, the final personalized group photo.

3.2 CanvasComposer Pipeline

CanvasComposer builds on a pretrained latent-based diffusion transformer (DiT) [Peebles and Xie 2023], as illustrated in Fig. 4. To

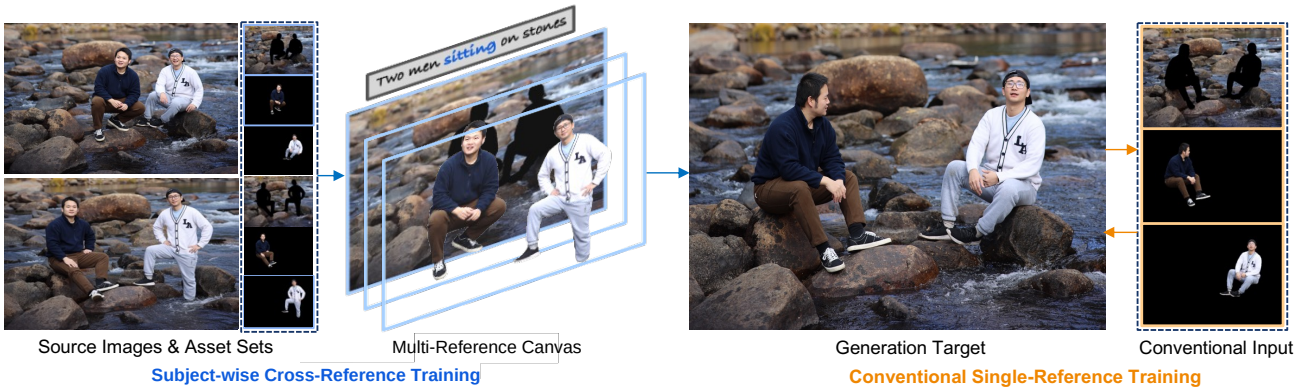

Fig. 3. **Subject-wise Cross-Reference Sampling Strategy.** During training, CanvasComposer constructs the multi-reference canvas by sampling each subject from a random source image within the same identity group, while the target image is a different sample from that group. This strategy intentionally introduces mismatches both among the input images and between the inputs and the target, encouraging CanvasComposer to learn adaptive alignment to the prompt and context during inference. In contrast, conventional single-reference training can overfit to pixel-level correspondences, often leading to outpainting or copy-paste artifacts (see examples in Fig. 9).

preserve the pretrained model's capacity as much as possible, CanvasComposer treats the multi-reference canvas as pruned in-context visual tokens (reference tokens processed jointly with the noisy image tokens) and reuses the DiT's native self-attention and 3D Rotary Position Embedding (RoPE) [Su et al. 2024] interface for element-aware conditioning, without adding a decoupled attention branch or per-subject adapter.

Per-Element Latent Extraction. For each input element $l_i \in L$, we first encode the RGB content using the pretrained VAE encoder to obtain its latents $z_i = \mathcal{E}(l_i^{\text{RGB}}) \in \mathbb{R}^{H' \times W' \times D}$ where $\mathcal{E}$ is the VAE encoder, H' and W' are the height and width of the latent grid, and D is the feature dimension.

Positional Embeddings for Canvas Elements. We adopt a simple yet effective positional embedding scheme to encode both the 2D latent-grid coordinate and which element each latent token belongs to, following the 3D RoPE parametrization of FLUX Kontext [Black Forest Labs 2025]. For every element's latent z_i, we construct a 3D positional embedding

$$\text{pos}_i = [j_i, x, y] \in \mathbb{R}^3, \tag{1}$$

where (x, y) are the coordinates in the latent grid and $j_i \in \{0, 1, \ldots, N\}$ indexes which canvas element a token belongs to. FLUX Kontext already uses this 3D RoPE interface; we reuse its parametrization unchanged (full formulation in *Appendix*). The original model sets $j_i = 0$ for image tokens; we reserve this index for the noisy latent tokens of the base diffusion model and assign $j_i \geq 1$ to each user-placed element on the canvas. This design preserves the pretrained positional interface of the base model while allowing attention layers to tell overlapping canvas elements apart. We emphasize that the index j_i acts purely as an *input*-side identity-separation signal: it follows the segmenter's confidence-ranked mask order and carries no depth or z-ordering semantics. Accordingly, CanvasComposer resolves occlusion within the *input* canvas by preserving each reference's pixels when subjects overlap (Tab. 3); explicit control over occlusion or depth ordering in the *output* image is left as future work, e.g. by integrating training-free occlusion control [Zhan and Liu 2025] or by training with depth-conditioned inputs, which would require depth estimation during data generation. Such positional embeddings are injected into the attention blocks through RoPE [Su et al. 2024], following the design of the base model.

Transparent Latent Pruning. To improve the efficiency of multi-subject personalization, we introduce the transparent latent pruning strategy that selectively retains the latent tokens from valid non-transparent regions according to the alpha channel, while discarding the rest. Concretely, for each element's alpha channel l_i^{α}, we first downsample it to the latent resolution:

$$\alpha_i^{\text{latent}} = \text{NearestResize}(l_i^{\alpha}) \in \mathbb{R}^{H' \times W'}. \tag{2}$$

We then apply alpha-based masking to the latent tokens z_i, keeping only those in regions with non-zero alpha values:

$$z_i^{\text{valid}} = \text{Concat}(\{z_i(x, y) | \alpha_i^{\text{latent}}(x, y) > 0.5\}), \tag{3}$$

where $z_i(x, y)$ and $\alpha_i^{\text{latent}}(x, y)$ denote the latent tokens and their alpha values at latent-grid coordinates (x, y).

Training and inference use the identical token-selection rule. Transparent RGB regions are not used as explicit conditioning tokens: after VAE encoding, tokens whose downsampled alpha is transparent are discarded before the DiT receives the conditioning sequence. Although VAE tokens near boundaries may contain local context from neighboring RGB pixels, the model is fine-tuned with this exact masking formulation, and only retained tokens are exposed to the DiT. Practically, we see no boundary artifacts or identity leakage from transparent regions, so explicit partial or masked convolutions are unnecessary for our setting.

Unlike existing personalization methods [Qian et al. 2025a; Zhang et al. 2025b] that allocate a fixed-length token block per subject on a full-canvas crop and thus scale linearly with the subject count, *the transparent latent pruning strategy makes the length of the token sequence proportional to the non-transparent content placed on the shared canvas, rather than to the number of subjects placed on the*

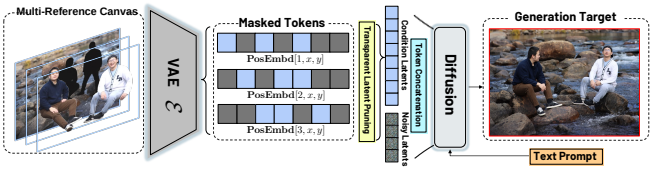

Fig. 4. **CanvasComposer Pipeline.** CanvasComposer conditions a diffusion model on a text prompt and a multi-reference canvas. Each RGBA image is encoded into latent tokens with a VAE and assigned a unique index j, yielding 3D positional embeddings $[j, x, y]$ that tell overlapping subjects apart. Transparent latent pruning keeps only non-transparent conditioning tokens and discards transparent regions (gray boxes), reducing the conditioning sequence for efficient personalized generation.

canvas. Because users place and resize segmented subjects on a single fixed canvas, per-subject footprints shrink as more subjects are added, thus yielding substantial efficiency improvements.

Combining the Conditioning Tokens. Finally, we construct the conditional latents by aggregating the pruned latents from every element on the canvas: $z_{\text{cond}} = \text{Concat}(z_1^{\text{valid}}, z_2^{\text{valid}}, \ldots, z_N^{\text{valid}})$, which is further concatenated with the noisy image latents z_t to form the latent input of the DiT model, as shown in Fig. 4.

3.3 CanvasComposer Training

Subject-wise Cross-Reference Data Sampling Strategy. CanvasComposer training employs a *subject-wise cross-reference* data sampling strategy, as illustrated in Fig. 3, to mitigate copy-paste artifacts. Our training requires a multi-image-per-scene dataset (Sec. 4.1), where each scene contains several images depicting the same human subjects under different poses, viewpoints, and lighting conditions. For each training sample, we randomly select an image as the ground-truth target, denoted I^{target}, and use the remaining images in the same scene as a pool of cross-reference source images. For every subject i in the target, we collect an asset set $\mathcal{A}_i$ consisting of its segments across the remaining source images within the scene.

To construct the input multi-reference canvas L, we build it one subject at a time. For each subject i, we randomly sample one asset $a_i \in \mathcal{A}_i$ and fit it to the subject's bounding box in I^{target}, giving a distinct RGBA image l_i. This training canvas uses target boxes only for supervised data construction, while sourcing appearance cues for each subject from different images within the same scene.

This subject-wise cross-reference construction produces training inputs where different subjects originate from different source images, rather than from a single reference frame as in the conventional training paradigm (see Fig. 3 right). Our strategy encourages CanvasComposer to mitigate copy-paste artifacts by integrating cross-image identity cues and to generalize across diverse poses, expressions, viewpoints, and contexts.

Canvas-Conditioned Finetuning. We adapt the pretrained DiT with lightweight LoRA finetuning using a flow matching loss [Lipman et al. 2023] with multi-reference canvas conditioning:

$$\mathcal{L}_{\text{cond}} = \mathbb{E}_{\substack{t\sim\mathcal{U}(0,1),z_0,\\ z_1,z_{\text{cond}},P}}\left[\left\|v_\theta(z_t,t,z_{\text{cond}},P)-(z_1-z_0)\right\|^2\right] \quad (4)$$

where $v_\theta(\cdot)$ is the predicted velocity, z_1 and z_0 are the latents of the target image and the sampled noise, z_t is the noisy latents at timestep t of the target image I^{target}, z_{cond} is the conditional latents of our multi-reference canvas L, and P is the text prompt, respectively.

4 EXPERIMENTS

4.1 Experimental Setup

Training Dataset Curation. Our training set comprises 32M in-house images across 6M scenes, with a focus on human subjects[1]. The training images are paired with standard captions, which provide the text prompt used to condition global scene content during training. We filter the data to ensure each scene contains at most 4 identities and to exclude low-resolution, low-quality faces. To construct our multi-reference training data, we apply an in-house, YOLO11-style [Khanam and Hussain 2024] human instance segmenter (a public YOLO11-seg model can serve as a substitute) to extract each human as a separate RGBA cutout and retain the remaining pixels as background. When constructing the input multi-reference canvas in each training step, we apply per-subject augmentations including random geometric jitter, color jitter with brightness ±0.5, contrast in [0.5, 1.0], saturation ±0.5, no hue shift, and object-level Gaussian blur and grayscale augmentation with probability 0.1 each.

Training Details. We train a LoRA with a rank of 512 on all attention layers of the frozen FLUX Kontext [Black Forest Labs 2025] using the AdamW optimizer [Loshchilov and Hutter 2019]. The original attention projection matrices are 3072×3072, making rank 512 a compact but expressive adaptation. The model is trained for 200K iterations with a constant learning rate of 1×10^{-4}, a total batch size of 32, and a 512×512 resolution. The entire training took 4 days on 4 nodes, each with 8 A100 GPUs.

Evaluation Details. We evaluate at 1024×1024 resolution using 128 images from FFHQ-in-the-wild [Karras et al. 2019] as identity inputs. FFHQ is a public, single-frame dataset and is not included in training. There are 32 prompts for each benchmark. All evaluations are conducted with 28 denoising steps for CanvasComposer, without any post-processing. See *Appendix* for evaluation details.

4.2 Baseline Comparisons

Four-Person (4P) Personalization. Most existing human-centric personalization methods [Mou et al. 2025; Qian et al. 2025a; Ye et al. 2023; Zhou et al. 2024] struggle in four-person settings because their computation and memory grow with the number of subjects. In contrast, CanvasComposer, enabled by our multi-reference canvas and transparent latent pruning, supports multiple subjects with substantially lower conditioning cost.

We compare CanvasComposer in the 4P setting against recent multi-subject personalization methods: UniPortrait [He et al. 2025], ID-Patch [Zhang et al. 2025b], UNO [Wu et al. 2025a], and OmniGen2 [Wu et al. 2025c]. CanvasComposer demonstrates stronger

[1]Our internal dataset cannot be released due to internal data policies. Following [Qian et al. 2025b], a similar dataset can be curated by sampling from public videos.

Fig. 5. **Qualitative Comparison in Four-person (4P) Personalization.** State-of-the-art baselines often fail to preserve all reference identities, causing identity swaps, distortions, omissions, or duplicated subjects. CanvasComposer consistently generates high-fidelity results that preserve all identities in challenging multi-person scenes. The "Inputs" column shows the multi-reference-canvas visualization for CanvasComposer, whereas all baselines receive individual images as required by their input formats throughout all benchmarks.

performance in this task. As shown in Fig. 5, CanvasComposer generates high-quality images that preserve all reference identities in challenging multi-person scenes.

Quantitatively, as reported in Tab. 1, CanvasComposer achieves the highest identity preservation by a large margin while maintaining competitive text-image alignment and aesthetic quality, as assessed by ArcFace [Deng et al. 2022], VQAScore [Lin et al. 2024], and HPSv3 [Ma et al. 2025], respectively. Our user study reveals a strong preference towards CanvasComposer, where participants favored it over all state-of-the-art 4P personalization baselines in over 96% of pairwise comparisons on average. We further compare against recent image-editing approaches, FLUX Kontext [Black Forest Labs 2025], Overlay Kontext [ilkerzgi and gokaygokay 2025], Qwen-Image-Edit [Wu et al. 2025b], and Nano-Banana [Comanici et al. 2025], on this challenging 4P task, as these methods have recently emerged as strong baselines for multi-subject personalization. As shown in Tab. 1, CanvasComposer outperforms all of them, achieving the highest identity preservation and the strongest user preference (see *Appendix* for qualitative comparisons and details).

We note an input asymmetry in these comparisons: CanvasComposer and the editing baselines receive the composited canvas that encodes the intended layout, whereas prior personalization baselines accept only individual reference images without spatial layout control.

Two-Person (2P) Personalization. Most existing personalization methods are designed specifically for up to 2P personalization. Here, we evaluate CanvasComposer for the 2P personalization task against the aforementioned 4P personalization methods and recent 2P personalization approaches including StoryMaker [Zhou et al. 2024] and DreamO [Mou et al. 2025]. As illustrated in Fig. 6, previous approaches again frequently struggle in two-person generation. They may omit one subject, duplicate identities, or fail to retain facial identity information, resulting in unnatural interactions or low-quality outputs. In contrast, our method consistently produces high-fidelity outputs where both identities are faithfully preserved.

Fig. 6. **Qualitative Comparison in Two-person (2P) Personalization.** Competing methods can miss one identity, mix identities between subjects, or copy the references too rigidly. CanvasComposer produces coherent two-person results that preserve both reference identities.

Method	ArcFace ↑	HPSv3 ↑	VQAScore ↑	Ours Win Rate (%) ↑
4P Personalization				
UniPortrait [He et al. 2025]	0.309	12.4	0.786	95.6
ID-Patch [Zhang et al. 2025b]	0.082	7.13	0.785	98.0
UNO [Wu et al. 2025a]	0.077	12.2	0.840	94.3
OmniGen2 [Wu et al. 2025c]	0.086	**13.0**	0.805	97.8
FLUX Kontext [Black Forest Labs 2025]	0.217	12.8	0.869	92.2
Overlay Kontext [ilkerzgi and gokaygokay 2025]	0.251	11.2	0.828	93.0
Qwen-Image-Edit [Wu et al. 2025b]	0.223	**13.0**	**0.895**	68.9
Nano-Banana [Comanici et al. 2025]	0.434	10.4	0.826	60.6
CanvasComposer (Ours)	**0.533**	12.5	0.840	Ref.
2P Personalization				
UniPortrait [He et al. 2025]	0.460	**13.2**	0.812	86.9
StoryMaker [Zhou et al. 2024]	0.542	4.97	0.523	92.8
ID-Patch [Zhang et al. 2025b]	0.121	10.1	0.858	98.4
UNO [Wu et al. 2025a]	0.072	11.2	0.870	93.1
DreamO [Mou et al. 2025]	0.443	12.4	**0.877**	93.4
OmniGen2 [Wu et al. 2025c]	0.121	12.8	0.828	84.7
CanvasComposer (Ours)	**0.547**	11.6	0.865	Ref.

Table 1. **Quantitative comparison.** CanvasComposer consistently achieves superior identity preservation (ArcFace [Deng et al. 2022]) in the 4P and 2P setups while maintaining strong aesthetic quality (HPSv3 [Ma et al. 2025]) and prompt alignment (VQAScore [Lin et al. 2024]). We also conduct a pairwise user study detailed in *Appendix*, where "Ours Win Rate (%)" denotes the fraction of comparisons in which our method is preferred over a baseline. Such user studies are important for personalization evaluation as automatic metrics can be noisy or biased. Across the 4P and 2P benchmarks, CanvasComposer is preferred over every baseline: winning over 84% of comparisons against personalization methods and over 60% against recent image-editing models (FLUX Kontext, Overlay Kontext, Qwen-Image-Edit, and Nano-Banana).

These qualitative trends are corroborated by quantitative evaluations in Tab. 1. CanvasComposer is strongly preferred by users and achieves the best identity preservation among all baselines.

Fig. 8. **Effect of Multi-Reference Canvas Layout.** The multi-reference canvas defines the relative spatial arrangement of subjects while the text prompt controls pose, interaction, and scene context.

Single-Person (1P) Personalization. Although our primary focus is multi-subject personalization, CanvasComposer is equally effective for single-subject scenarios. As shown in Fig. 7, competing approaches tend to inject the reference identity with limited pose or expression variability, whereas CanvasComposer faithfully follows diverse prompts. See *Appendix* for the quantitative comparisons and the comparisons with single-subject personalization methods.

4.3 Ablation Study and Analysis

Our contributions primarily address interactive multi-reference placement, occlusion handling, and training robustness through subject-wise cross-reference training and the multi-reference canvas. Their effectiveness is demonstrated in Fig. 9. We include FLUX Kontext, the base model of CanvasComposer, in the ablation.

Effect of Subject-Wise Cross-Reference Training. Removing this component leads to copy-paste artifacts and degraded image quality, producing results that resemble naive outpainting rather than coherent multi-subject synthesis. In contrast, our full model preserves

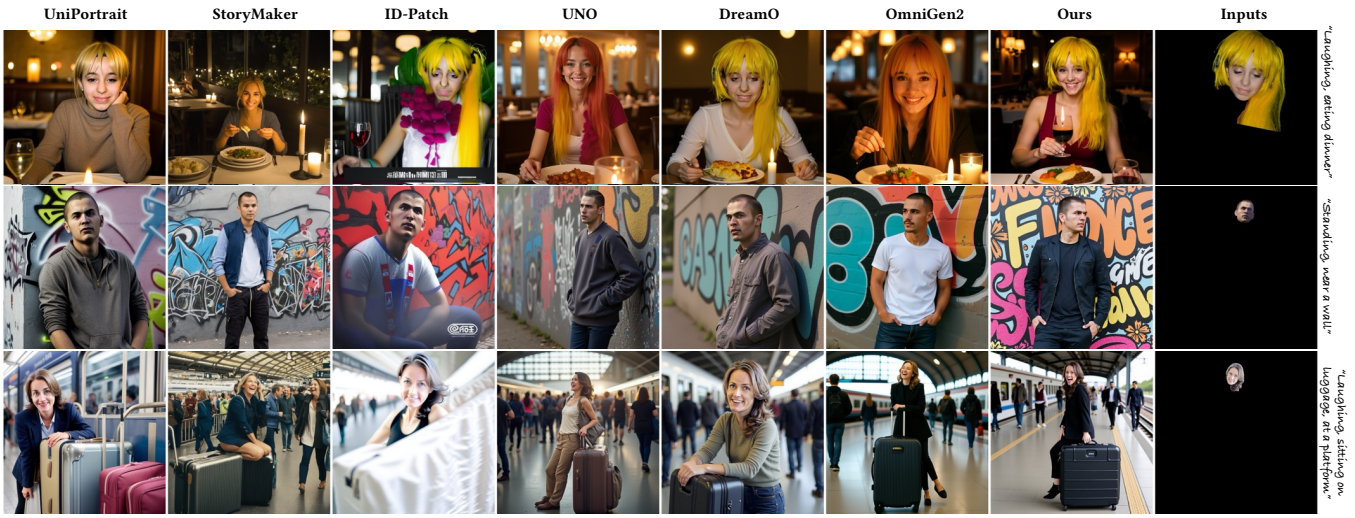

Fig. 7. **Qualitative Comparison in Single-Person (1P) Personalization.** State-of-the-art methods often fail to decouple identity from pose and expression, leading to copy-paste artifacts, whereas CanvasComposer remains faithful to the reference identity while following diverse prompts (e.g. *smiling* and *laughing*).

Method	#tokens ↓	Latency (seconds) ↓
Ours	2200.2	21.2
w/o Multi-Reference Canvas	5120	44.3
w/o Transparent Pruning	16384	91.8

Table 2. **Efficiency Analysis on the 4P Personalization Benchmark.**

	2P Personalization			2P Overlap		
Method	**Arc. ↑**	**HPS ↑**	**VQA ↑**	**Arc. ↑**	**HPS ↑**	**VQA ↑**
Ours	**0.547**	**11.6**	**0.865**	**0.416**	**11.4**	**0.764**
w/o Multi-Reference Canvas	0.525	11.5	0.846	0.164	11.28	0.752
IP-Adapter (Ours)	0.485	10.2	0.856		—	

Table 3. **Multi-reference canvas and adapter ablations.** "w/o Multi-Reference Canvas" is the collage-input baseline: the model receives a single composited RGB canvas instead of separate per-subject images. IP-Adapter (Ours) is trained on the same data as CanvasComposer under the same 2P evaluation protocol.

identities more naturally and avoids copying facial expressions from the inputs due to the disentanglement introduced by our training.

Effect of Multi-Reference Canvas. Without the multi-reference canvas, the model is trained on a single collage image ("Input" column in Fig. 9) without transparent latent pruning. As the number of personalized subjects increases, occlusion becomes inevitable, and such a single-canvas representation fails to preserve fine details or properly disentangle overlapping subjects. Occluded details in the collage, e.g., the pom-pom on the red hat of the left woman in the 1st row, are completely lost, and identity preservation for the occluded person degrades substantially, as shown in the second row. In contrast, CanvasComposer handles these scenarios robustly: keeping each subject spatially independent preserves each reference even when subjects overlap, preventing occlusion-induced artifacts. Fig. 8 further shows that the canvas layout controls the spatial arrangement of generated subjects.

Efficiency Analysis. Tab. 2 shows that transparent latent pruning reduces the average number of conditioning tokens by over 85% and lowers inference latency from 91.8s to 21.2s on the 4P Personalization Benchmark.

Overlap Benchmark. To isolate the benefit of keeping each subject spatially separate, our 2P Overlap Benchmark centers two references in the same canvas region, so the composited RGB input occludes the background reference while CanvasComposer retains both references. Tab. 3 shows that the variant without the multi-reference canvas is competitive on the standard 2P benchmark but loses substantial identity fidelity under overlap, while CanvasComposer remains much more robust; Fig. 11 visualizes such failure modes.

Adapter Baseline. Tab. 3 also compares CanvasComposer with a same-data, same-training IP-Adapter baseline on the 2P Personalization Benchmark; the multi-reference canvas provides stronger identity preservation and image quality than adapter-style conditioning.

Beyond-Human Subjects. Although our main experiments focus on human subjects, Fig. 10 shows anecdotal examples where a non-human object is placed on the canvas as an additional reference, suggesting a possible extension of the multi-reference representation beyond humans.

Personalization with Background. The multi-reference canvas can also accept an optional background image. As shown in Fig. 12, CanvasComposer generates images where humans interact naturally with the background, e.g. leaning against a tree trunk, under coherent lighting.

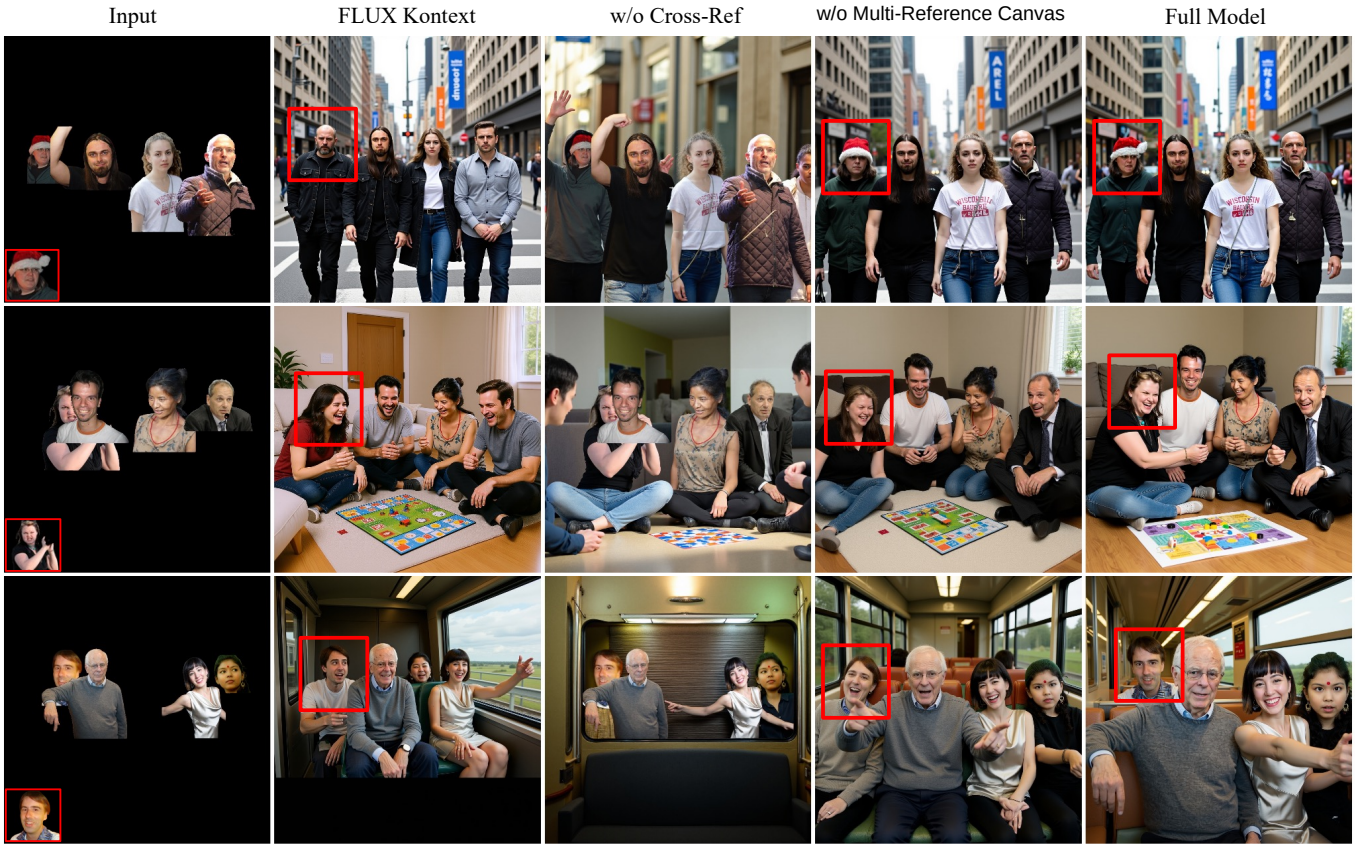

Fig. 9. **Ablation Study.** Cross-reference training improves multi-subject generation by mitigating copy-paste artifacts and encouraging subjects to adapt naturally to the scene. The multi-reference canvas addresses input occlusion (highlighted in red boxes): without keeping subjects on separate images, overlapping subject details (e.g., the pom-pom on the red hat of the left woman in the 1st row) disappear, and identity preservation deteriorates for occluded subjects.

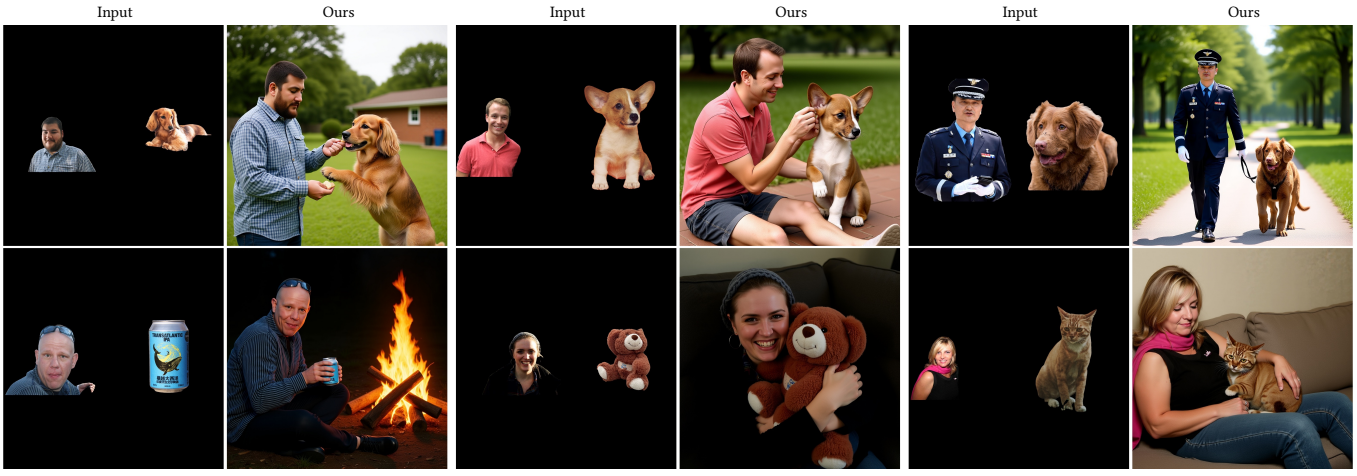

Fig. 10. **Results Beyond Humans.** Although CanvasComposer is trained only on human-centric data, it can accept non-human objects placed on the canvas at inference and produce coherent interactions between people and surrounding objects.

5 LIMITATIONS AND FUTURE WORK

Reasoning Limitations. The method sometimes struggles when the generated image requires detailed reasoning about how humans should interact with the background, e.g. failing to seat the

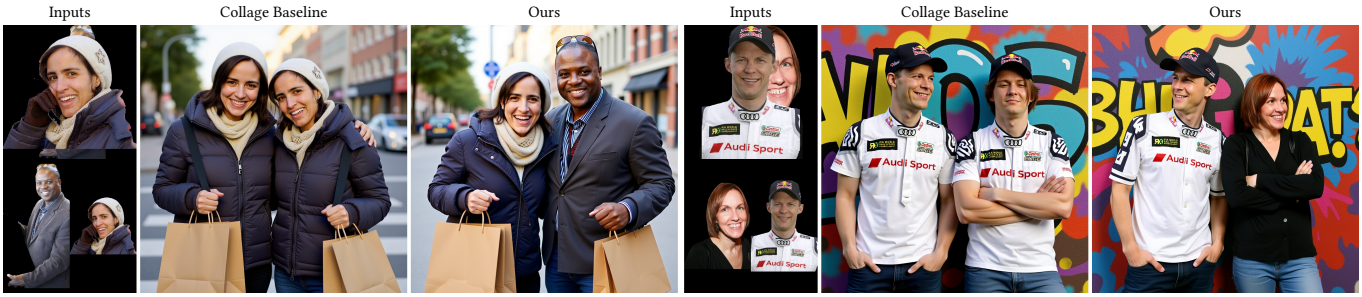

Fig. 11. **Qualitative Results on the 2P Overlap Benchmark.** In the Inputs column, the first row is the composited visualization of the multi-reference canvas, and the second row shows the reference inputs side by side for demonstration purposes only. The collage baseline (w/o Multi-Reference Canvas) loses the occluded reference while CanvasComposer preserves both identities.

Fig. 12. **4P Personalization with Background.** The multi-reference canvas can include an optional background, resulting in five separate images: four people and one background. The inserted humans interact naturally with the background, e.g. leaning against a tree trunk, under coherent lighting. Left: composited visualizations of the multi-reference canvas.

Benchmark	ArcFace ↑
4P	0.533
5P	0.206
6P	0.152

Table 4. **Beyond-4P Personalization.** Performance degrades beyond 4P.

#People in Prompt	#People in Generation
2P	2.25
4P	4.06
5P	7.75
6P	7.75

Table 5. **Base model beyond four people.** FLUX Kontext does not reliably follow prompts with more than four people.

foreground humans naturally on chairs (see the failure example in *Appendix*). Integrating the strong reasoning capabilities of Vision Language Models [Bai et al. 2023, 2025] into personalized generation is a promising remedy.

Beyond 4P Generation Limitations. Although CanvasComposer can accept a variable number of subjects, its performance degrades beyond four people (Tab. 4) for two reasons. First, **data limitations**: our >4P in-house samples often contain highly similar poses, expressions, or low-quality faces, so we cap the training data at four people; higher-quality, more diverse >4P data would likely improve robustness. Second, **base model limitations**: FLUX Kontext itself is unreliable beyond four subjects (Tab. 5); access to its raw base weights, prior to aesthetic fine-tuning or guidance distillation (as explored for FLUX.1-Krea [Labs 2025]), would likely enable further improvements.

Output Layout and Occlusion Control. The index CanvasComposer uses to keep subjects separate on the input canvas carries no depth or z-ordering semantics: CanvasComposer preserves reference pixels when subjects overlap in the *input* canvas (Tab. 3), while the occlusion and depth ordering of the *output* follow the base model. See Sec. 3.2 for future directions toward such controls.

Base Model Dependence. All results use a frozen FLUX Kontext base with a lightweight LoRA; performance inherits this backbone's strengths and weaknesses (e.g. beyond four people, Tabs. 4 and 5) and may vary on other base models.

Single Flattened Output. CanvasComposer targets a single harmonized multi-human group photo and does not produce layered or transparent outputs. As a natural extension, combining our input canvas with transparent-image generation techniques [Fontanella et al. 2024; Li et al. 2023a; Zhang and Agrawala 2024] to produce layered RGBA outputs is compelling future work.

6 CONCLUSION

In this paper, we introduced CanvasComposer, an interactive framework for multi-human personalized generation. Users can directly place and resize multiple input references on the canvas: by representing each reference subject as its own separate image, CanvasComposer supports intuitive layout manipulation while preserving reference content under occlusion. The references are kept distinct simply by repurposing the RoPE embeddings of the pretrained base model, assigning each reference its own index. Transparent latent pruning further reduces unnecessary conditioning tokens, and cross-reference training mitigates copy-paste artifacts. Experiments across multi-human benchmarks show that CanvasComposer improves identity preservation and generation quality over existing methods. More broadly, the multi-reference canvas is an effective interface for adapting pretrained diffusion models to multi-human personalization. We believe CanvasComposer opens several promising directions for future work.

REFERENCES

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia*

2023 Conference Papers. 1–12.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR* abs/2308.12966 (2023).

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

Black Forest Labs. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742* (2025).

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. 2025. Multi-subject open-set personalization in video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6099–6110.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2024. AnyDoor: Zero-shot Object-level Image Customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6593–6602.

Bo Cheng, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. 2024. HiCo: hierarchical controllable diffusion model for layout-to-image generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*. 128886–128910.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. 2024. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*. Springer, 432–448.

Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. 2025. LoRAShop: Training-Free Multi-Concept Image Generation and Editing with Rectified Flow Transformers. *CoRR* abs/2505.23758 (2025).

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. 2022. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 10 (2022), 5962–5979.

Alessandro Fontanella, Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, and Sarah Parisot. 2024. Generating Compositional Scenes via Text-to-Image RGBA Instance Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*. OpenReview.net.

Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2024. LCM-Lookahead for Encoder-Based Text-to-Image Personalization. In *ECCV (14) (Lecture Notes in Computer Science, Vol. 15072)*. Springer, 322–340.

Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. 2025. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions on Graphics (TOG)* 44, 4 (2025), 1–11.

Anujraaj Argo Goyal, Guocheng Gordon Qian, Huseyin Coskun, Aarush Gupta, Himmy Tam, Daniil Ostashev, Ju Hu, Dhritiman Sagar, Sergey Tulyakov, Kfir Aberman, and Kuan-Chieh Jackson Wang. 2025. Preventing Shortcuts in Adapter Training via Providing the Shortcuts. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. 2024. PuLID: Pure and Lightning ID Customization via Contrastive Alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yucheng Han, Rui Wang, Chi Zhang, Juntao Hu, Pei Cheng, Bin Fu, and Hanwang Zhang. 2024. EMMA: Your Text-to-Image Diffusion Model Can Secretly Accept Multi-Modal Prompts. *CoRR* abs/2406.09162 (2024).

Junjie He, Yifeng Geng, and Liefeng Bo. 2025. UniPortrait: A Unified Framework for Identity-Preserving Single- and Multi-Human Image Personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

ilkerzgi and gokaygokay. 2025. Overlay-Kontext-Dev-LoRA. https://huggingface.co/ilkerzgi/Overlay-Kontext-Dev-LoRA. LoRA fine-tune of FLUX.1-Kontext-dev for image overlay tasks.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*. 4401–4410.

Rahima Khanam and Muhammad Hussain. 2024. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv preprint arXiv:2410.17725* (2024).

Chanran Kim, Jeongin Lee, Shichang Joung, Bongmo Kim, and Yeul-Min Baek. 2024. InstantFamily: Masked Attention for Zero-shot Multi-ID Image Generation. *CoRR* abs/2404.19427 (2024).

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. 2024. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *European Conference on Computer Vision*. Springer, 253–270.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *CVPR*. 1931–1941.

Black Forest Labs. 2025. FLUX.1-Krea-dev. https://huggingface.co/black-forest-labs/FLUX.1-Krea-dev.

Phillip Y. Lee, Taehoon Yoon, and Minhyuk Sung. 2024. GrounDiT: Grounding Diffusion Transformers via Noisy Patch Transplantation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pengzhi Li, Qinxuan Huang, Yikang Ding, and Zhiheng Li. 2023a. Layerdiffusion: Layered controlled image editing with diffusion models. In *SIGGRAPH Asia 2023 Technical Communications*. 1–4.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22511–22521.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2024. PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 8640–8650.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow Matching for Generative Modeling. In *ICLR*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. 2025. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhenxing Mi, Kuan-Chieh Wang, Guocheng Qian, Hanrong Ye, Runtao Liu, Sergey Tulyakov, Kfir Aberman, and Dan Xu. 2025. I Think, Therefore I Diffuse: Enabling Multimodal In-Context Reasoning in Diffusion Models. *ICML* (2025).

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. 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*, Vol. 38. 4296–4304.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, Mengtian Li, Songtao Zhao, Jian Zhang, Qian He, and Xinglong Wu. 2025. DreamO: A Unified Framework for Image Customization. In *SIGGRAPH Asia 2025 Conference Papers*.

Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. 2022. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–10.

Yotam Nitzan, Zongze Wu, Richard Zhang, Eli Shechtman, Daniel Cohen-Or, Taesung Park, and Michaël Gharbi. 2024. Lazy Diffusion Transformer for Interactive Image Editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 55–72.

Or Patashnik, Rinon Gal, Daniil Ostashev, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. 2025. Nested Attention: Semantic-aware Attention Values for Concept Personalization. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–12.

William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 4172–4182.

Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. 2024. Orthogonal Adaptation for Modular Customization of Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 7964–7973.

Guocheng Qian, Kuan-Chieh Wang, Or Patashnik, Negin Heravi, Daniil Ostashev, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. 2025b. Omni-ID: Holistic Identity Representation Designed for Generative Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8786–8795.

Guocheng Gordon Qian, Daniil Ostashev, Egor Nemchinov, Avihay Assouline, Sergey Tulyakov, Kuan-Chieh Jackson Wang, and Kfir Aberman. 2025a. ComposeMe: Attribute-Specific Image Prompts for Controllable Human Image Generation. In *SIGGRAPH Asia 2025 Conference Papers*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Jiawei Ren, Mengmeng Xu, Jui-Chieh Wu, Ziwei Liu, Tao Xiang, and Antoine Toisoul. 2024. Move anything with layered scene diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6380–6389.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*. 22500–22510.

Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. 2024. Collage diffusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4208–4217.

Takahiro Shirakawa and Seiichi Uchida. 2024. NoiseCollage: A Layout-Aware Text-to-Image Diffusion Model Based on Noise Cropping and Merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. 2023. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18310–18319.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.

Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. 2024b. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *SIGGRAPH Asia 2024 Conference Papers*. 1–12.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. 2024a. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519* (2024).

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. 2025. MS-Diffusion: Multi-subject Zero-shot Image Personalization with Layout Guidance. In *ICLR*. OpenReview.net.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025b. Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025).

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. 2025c. OmniGen2: Exploration to Advanced Multimodal Generation. *arXiv preprint arXiv:2506.18871* (2025).

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. 2025a. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. 2025b. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision* 133, 3 (2025), 1175–1194.

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

Zhexiao Xiong, Wei Xiong, Jing Shi, He Zhang, Yizhi Song, and Nathan Jacobs. 2024. Groundingbooth: Grounding text-to-image customization. *arXiv preprint arXiv:2409.08520*.

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

Xiaohang Zhan and Dingming Liu. 2025. LaRender: Training-Free Occlusion Control in Image Generation via Latent Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Hui Zhang, Dexiang Hong, Maoke Yang, Yutao Cheng, Zhao Zhang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. 2025a. Creatidesign: A unified multi-conditional diffusion transformer for creative graphic design. *arXiv preprint arXiv:2505.19114* (2025).

Lvmin Zhang and Maneesh Agrawala. 2024. Transparent Image Layer Diffusion using Latent Transparency. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–15.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Yimeng Zhang, Tiancheng Zhi, Jing Liu, Shen Sang, Liming Jiang, Qing Yan, Sijia Liu, and Linjie Luo. 2025b. ID-Patch: Robust ID Association for Group Photo Personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2986–2996.

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

Zhengguang Zhou, Jing Li, Huaxia Li, Nemo Chen, and Xu Tang. 2024. Storymaker: Towards holistic consistent characters in text-to-image generation. *arXiv preprint arXiv:2409.12576* (2024).