Canvas-to-Image: Compositional Image Generation with Multimodal Controls

Yusuf Dalva[*1,3]   Guocheng Gordon Qian[*†1]   Maya Goldenberg[1]   Tsai-Shien Chen[1,2]
Kfir Aberman[1]   Sergey Tulyakov[1]   Pinar Yanardag[3]   Kuan-Chieh Jackson Wang[1]

[1]Snap Inc.   [2]UC Merced   [3]Virginia Tech
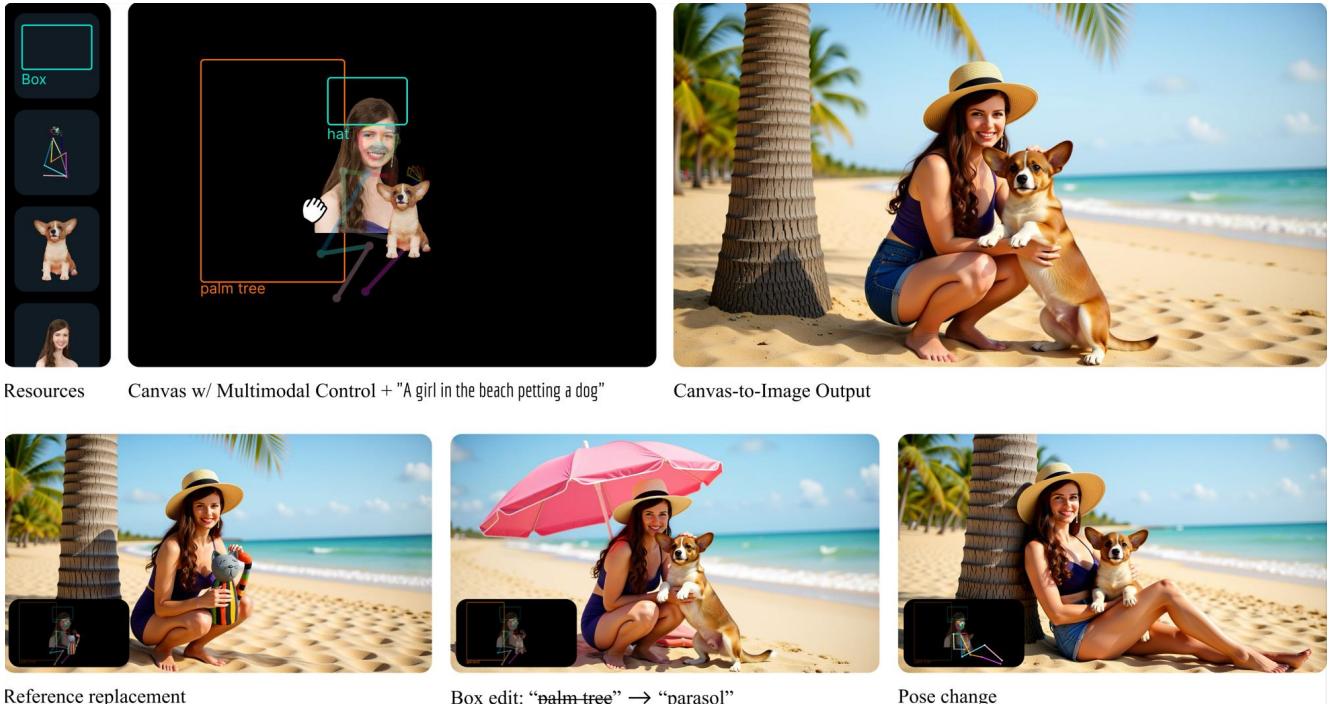https://snap-research.github.io/canvas-to-image/

Figure 1. **Canvas-to-Image** enables compositional control for text-to-image generation through a unified Multi-Task Canvas framework. The canvas serves as a flexible visual interface that guides image synthesis by supporting diverse guiding signals, including spatially positioned subjects, pose signals, bounding boxes, and text annotations.

Abstract

While modern diffusion models excel at generating high-quality and diverse images, they still struggle with high-fidelity compositional and multimodal control, particularly when users simultaneously specify text prompts, subject references, spatial arrangements, pose constraints, and layout annotations. We introduce Canvas-to-Image, a unified framework that consolidates these heterogeneous controls into a single canvas interface, enabling users to generate images that faithfully reflect their intent. Our key idea is to encode diverse control signals into a single compos-ite canvas image that the model can directly interpret for integrated visual-spatial reasoning. We further curate a suite of multi-task datasets and propose a Multi-Task Canvas Training strategy that optimizes the diffusion model to jointly understand and integrate heterogeneous controls into text-to-image generation within a unified learning paradigm. This joint training enables Canvas-to-Image to reason across multiple control modalities rather than relying on task-specific heuristics, and it generalizes well to multi-control scenarios during inference. Extensive experiments show that Canvas-to-Image significantly outperforms state-of-the-art methods in identity preservation and control adherence across challenging benchmarks, including multi-

*Equal Contributions. †Corresponding author.

1

person composition, pose-controlled composition, layout-constrained generation, and multi-control generation.

1. Introduction

Recent advances in large-scale diffusion models [9, 33, 38] have substantially improved the realism and diversity of synthesized imagery. However, these models remain inherently stochastic and provide limited flexibility when users wish to control multiple aspects of image generation simultaneously. This limitation is particularly consequential in creative and design-oriented applications, such as digital art and content creation, where users often need to coordinate several types of control signals, such as spatial layouts, subject references, pose constraints, etc.

We introduce **Canvas-to-Image**, a framework that enables *heterogeneous compositional control* over diffusion-based image generation through a unified canvas representation. As illustrated in Fig. 1, our approach allows users to combine diverse forms of input within a single interface: subjects and objects can be positioned, resized, rotated, and posed; bounding boxes with descriptive tags can define subjects with spatial constraints; and pose overlays [4] can specify body configurations. This flexible, multi-modal interaction design enables users to guide the generation process using complementary controls that collaboratively define both semantics and composition.

Achieving unified multi-control generation remains highly challenging, and no existing model can handle all the aforementioned controls simultaneously. Existing control mechanisms [22, 26, 52] typically address *isolated* aspects of compositional image synthesis, e.g. spatial layouts or pose constraints, but *fail to handle multiple controls within a single input*. The core difficulty lies in reconciling heterogeneous inputs that differ in both structure and semantics, including subject references, bounding boxes, and textual tags, while training a model capable of jointly interpreting and balancing these signals.

Consequently, prior works supporting subject injection [14, 35, 50] usually lack spatial control, whereas layout-guided methods [22, 51] cannot incorporate specific poses or subjects. Recent methods such as StoryMaker [55] and ID-Patch [53] demonstrate both subject insertion and spatial control, but rely on complex module combinations, such as ControlNet [49] and IP-Adapter [50], which introduce additional complexity, are limited to face injection, lack bounding-box support, and generalize poorly.

To address these challenges, we propose three key innovations. **First**, we introduce the **Multi-Task Canvas**, a unified input representation that consolidates diverse control modalities, including background composition, subject insertion, bounding-box layouts, and pose guidance, into a single composite RGB image. This *canvas* serves as a gen-
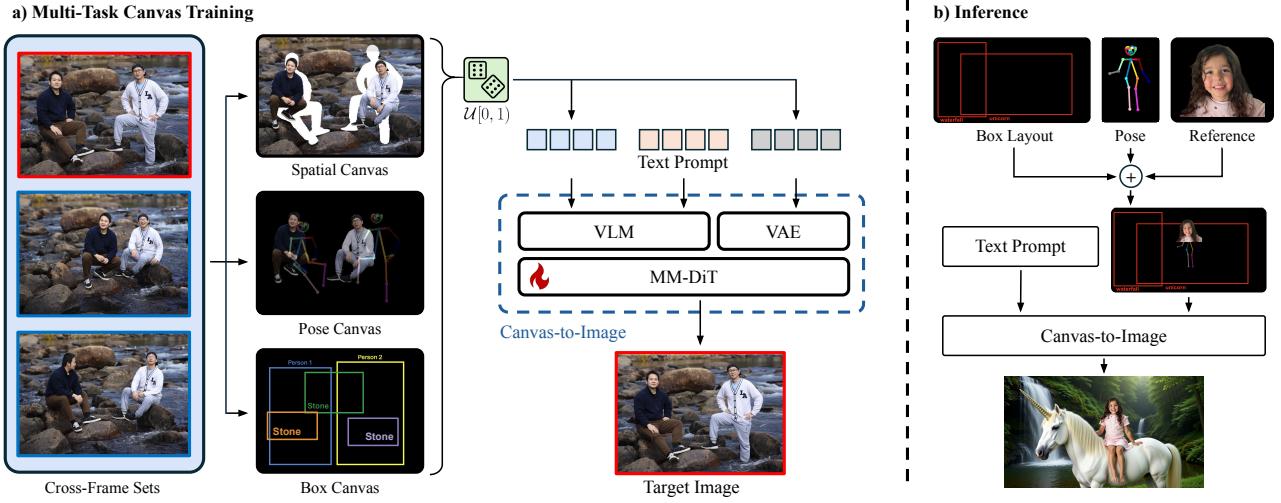
eralized visual interface where all control elements are expressed in a common pixel space, allowing the model to interpret multimodal guidance without extra modules or architectural changes. **Second**, we curate a comprehensive multi-task dataset that aligns these heterogeneous controls with corresponding target images, supporting consistent joint supervision across tasks. **Third**, we design a **Multi-Task Canvas Training** framework that fine-tunes the diffusion model to reason across tasks collectively, learning shared semantics and spatial dependencies among different control types. Importantly, we observe that once trained on this multi-task canvas, the model generalizes naturally to *multi-control* scenarios at inference time, even when combinations of controls were not seen together during training. We summarize our **contributions** as follows:

- **Unified canvas framework:** A generalized *Multi-Task Canvas* representation that consolidates heterogeneous controls into a single canvas-to-image formulation (Fig. 2), enabling coherent reasoning across modalities.
- **Multi-task datasets and training:** We curate comprehensive multi-task datasets covering diverse control modalities and propose a unified *Multi-Task Canvas Training* framework that fine-tunes the diffusion model jointly across these tasks. Experiments reveal joint training enables mixed controls in inference time.
- **Comprehensive evaluation:** Extensive experiments on challenging benchmarks demonstrate clear improvements in identity preservation and control adherence compared to existing methods. Ablations confirm that our unified multi-task design is key to achieving flexible and coherent heterogeneous control.

2. Related Work

Diffusion Models for Image Synthesis. Diffusion models [16, 41] are the dominant paradigm for high-fidelity image synthesis. Text-to-image models [37, 38, 40] use large-scale text-image pairs for open-vocabulary generation. Diffusion transformers [3, 9, 31] have further improved quality and scalability. Emerging multimodal models [7, 45] integrate MLLMs with diffusion models for higher quality and better prompt following. While impressive, these models still struggle with fine-grained, multi-constraint compositional control. Our work builds on a pretrained diffusion model, introducing a unified canvas interface and multi-task training strategy to enable comprehensive compositional control.

Personalization in Image Generation. Personalization methods generate specific subjects or identities in novel contexts. Early approaches [10, 21, 39] require per-concept fine-tuning. Adapter-based solutions [12, 13, 30, 34, 44, 50] improve efficiency by keeping the base model frozen and injecting subject-specific representations. Multi-concept personalization remains challenging: optimization-based

Figure 2. **Overview of Canvas-to-Image framework.** (a) **Multi-Task Canvas Training.** We reformulate heterogeneous control tasks: spatial composition, pose guidance, and layout-constrained generation into a single *canvas-to-image* formulation. Each training step samples one type of canvas (Spatial, Pose, or Box), where the target frame serves as supervision. All control signals are encoded as RGB canvases interpretable by the Vision-Language Model (VLM) for unified visual–spatial reasoning. The Multi-Modal DiT (MM-DiT) receives VLM embeddings, VAE latents, and noisy latents to predict the velocity for flow matching. (b) **Inference.** Although trained on single-control samples, the model generalizes to multi-control compositions, jointly leveraging pose, layout, and reference cues within a single generation process. This enables coherent multi-control reasoning without task-specific retraining.

methods [1, 6, 11, 20, 32] demand explicit concept disentanglement, while optimization-free approaches [5, 14, 35, 43, 48] concatenate embeddings at the cost of linear complexity growth. Beyond these scalability and flexibility limitations, most personalization methods focus solely on reference injection [15, 36]. A true creative design process requires handling multiple controls simultaneously. Canvas-to-Image addresses the scalability and flexible control challenges with a unified single canvas that maintains a constant computation cost, providing a foundation for such a multi-control, multi-subject personalization framework.

Compositional Control in Generation. Providing fine-grained compositional control remains a challenge, as existing mechanisms typically address isolated tasks. For instance, models like ControlNet [52] and T2I-Adapter [26] use structural cues like pose skeletons or depth maps to specify body configurations. Another line of work targets spatial layout control. Methods such as GLIGEN [22], LayoutDiffusion [54], and CreatiDesign [51] finetune the generator to interpret bounding boxes or segmentation masks. Unifying these heterogeneous controls is highly challenging, particularly with identity constraints for personalization. Methods supporting subject injection often lack fine-grained spatial control, while layout-guided methods cannot incorporate specific poses or subject identities. Recent attempts at unification, such as StoryMaker [55] and ID-Patch [53], rely on complex combinations of separate

modules (e.g., ControlNet with IP-Adapter) and are limited to single-type control. Canvas-to-Image addresses this gap by reformulating diverse control types into a single "visual canvas". Instead of relying on task-specific heuristics, our unified canvas supports spatial layouts, pose guidance, and subject appearance injection within one coherent interface, enabling the model to reason across modalities collectively.

3. Methodology

Canvas-to-Image is a unified framework for multi-modal, compositionally controlled image synthesis. The model takes as input a generalized Multi-Task Canvas, which is a single RGB image used to encode heterogeneous user controls. These controls include subject identities for personalization, spatial layouts, human poses, or bounding boxes. The Multi-Task Canvas formulation (Sec. 3.1) enables the diffusion model to interpret these diverse control modalities, all unified within this single image format, within a consistent training setup. Each canvas variant teaches the model a different type of compositional reasoning, from using subject references for personalization to applying fine-grained structural guidance. The underlying VLM–Diffusion architecture and multi-task training strategy (Sec. 3.2) jointly optimize the model across all control types. This design enables Canvas-to-Image to generalize to multi-control scenarios at inference, combining conditions not seen together during training while maintaining precise,

3

controllable synthesis.

3.1. Multi-Task Canvas

Our core contribution is the introduction of a Multi-Task Canvas that generalizes different complex compositional tasks into a shared input format: a single RGB image. This "visual canvas" serves as a flexible, multi-modal format that unifies diverse compositional inputs. We generate our canvas variants, which serve as different ways of expressing a composition, from data sources appropriate for each task. These variants are designed to be interpreted as distinct control types. For example, a Spatial Canvas provides a literal, pixel-based composition, while a Pose Canvas provides an abstract, structural one. Canvas-to-Image is built upon three primary canvas variants:

Spatial Canvas. The first variant trains the model to render a complete scene based on an explicit composition, as depicted in Fig. 2 as "Spatial Canvas". This input canvas is a composite RGB image created by visually pasting segmented cutouts of subjects (e.g., $I_{\text{subject\_1}}, I_{\text{subject\_2}}$) at their desired locations on a masked background. This canvas is constructed using Cross-Frame Sets (Fig. 2 left), which allows for pairing subjects and backgrounds drawn in a cross-frame manner. This strategy is crucial for the methodology as it avoids the copy-pasting artifacts common in simpler composition methods. This canvas enables multi-subject personalization as a compositional control, where users can place and resize reference subjects to guide the generation.

Pose Canvas. This task enhances the Spatial Canvas by providing a strong visual constraint for articulation. We overlay a ground-truth pose skeleton (e.g., from [4]) onto the Spatial Canvas using a specific transparency factor, as shown in Fig. 2 as "Pose Canvas". This semi-transparent overlay is a key design choice: the pose skeleton remains clearly recognizable as a structural guide, while the visual identity from the underlying subject segments (when present) can still be recovered and interpreted by the model. In this canvas, the subject segments themselves are randomly dropped during training, i.e., there are cases with only poses in the empty canvas to guide the pose. This is designed to support pose control as an independent modality in inference, even without reference injection.

Box Canvas. This task trains the model to interpret explicit layout specifications through bounding boxes with textual annotations directly onto the canvas. Each box contains a textual identifier (e.g., "Person 1", "Person 2", "Stone" in Fig. 2) that specifies which subject should appear in that spatial region and their sizes. The person identifier is ordered from left to right. Such a "Box Canvas" supports simple spatial control with text annotations without reference images as in previous two canvas variants.

By training the model on these distinct, single-task canvas types, the framework learns a robust and generalizable

policy for each control. Interestingly, this enables the model to generalize beyond single-task learning, allowing for the simultaneous execution of these distinct control signals at inference time even in combinations not encountered during training, as shown in Fig. 2(b).

3.2. Model and Multi-Task Training

As illustrated in Fig. 2, Canvas-to-Image builds upon a VLM–Diffusion architecture. The Vision-Language Model (VLM) encodes the unified canvas into a tokenized representation. This representation is concatenated with the VAE latents of the canvas and provided to the diffusion model as conditional inputs, along with the text prompt embedding and the noisy latents. The model is optimized using a task-aware Flow-Matching Loss:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{\substack{x_0, x_1, t, \\ h, c}} \left[\left\| v_\theta \big(x_t, t, [h; c] \big) - (x_0 - x_1) \right\|_2^2 \right], \quad (1)$$
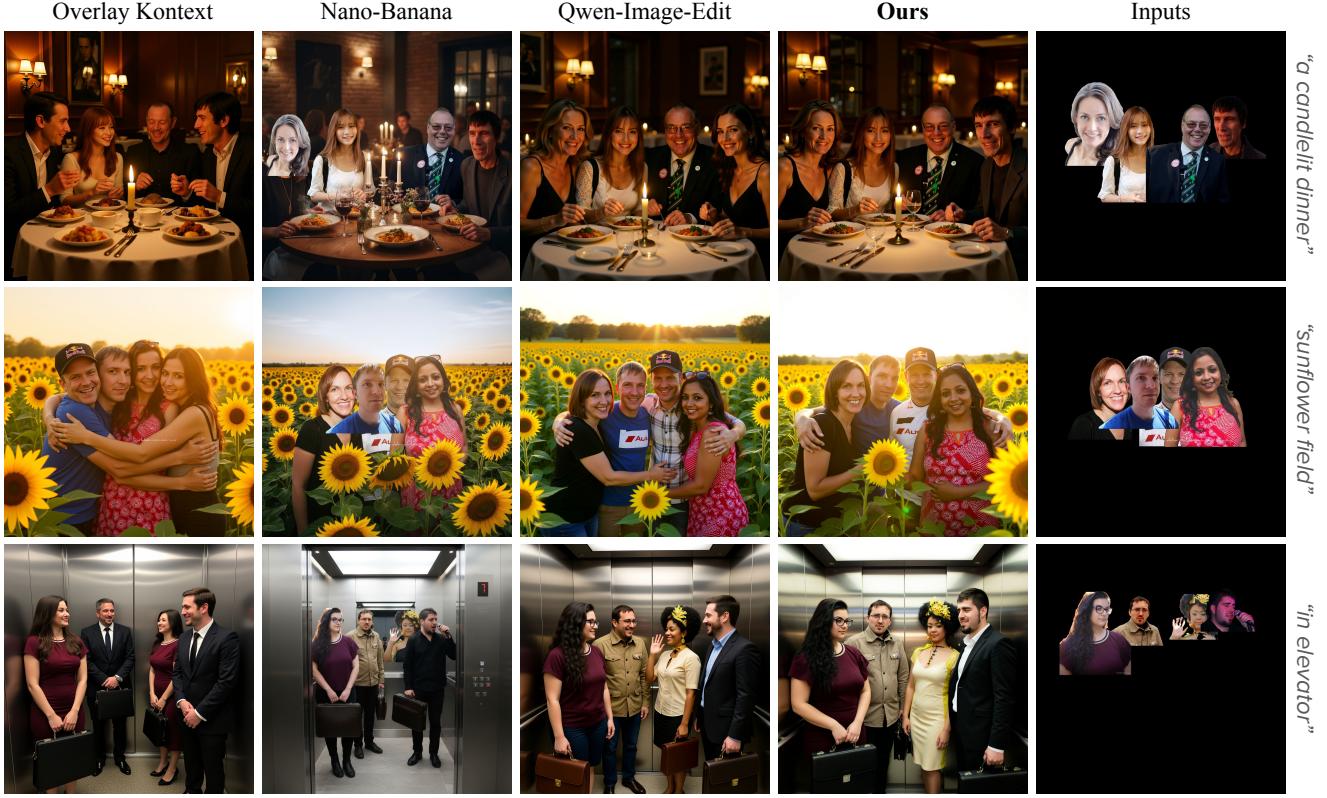
where x_0 is the target latent, x_1 is the noise latent, and x_t is the interpolated latent. h represents the input condition, which is itself a concatenation of the VLM embeddings (derived from the canvas and text prompt) and the VAE latents (derived from the same canvas). c represents the task indicator that specifies the current control type. The network v_θ predicts the target velocity $v_t = x_0 - x_1$.

Canvas-to-Image adopts a unified *Multi-Task Canvas* formulation, in which each training step samples one type of canvas as the input condition (*e.g.*, *Spatial*, *Pose*, *Box*). Training on this diverse, multi-task curriculum enables the model to learn decoupled, generalizable representations for each control type. Consequently, the model can execute a combination of these controls at inference time (e.g., a mixed canvas with both pose skeletons and layout boxes) despite having never seen such a combination during training. This emergent generalization from single-task learning to multi-task application is a key property of the proposed framework. To prevent task interference, we introduce a task indicator prompt—a short textual token (e.g., "[Spatial]", "[Pose]" or "[Box]") prepended to the user prompt. This indicator (c), which is necessary because our different canvas types represent different control meanings, disambiguates the task context and prevents mode blending. Ablation studies (Sec. 4.3) demonstrate the effectiveness of our multi-task training strategy on performing these control tasks compositionally at inference time.

4. Experiments

4.1. Experiment Details

Implementation. We build upon Qwen-Image-Edit [45] as our base architecture. The input canvas image and text prompt are first processed by the VLM to extract seman-

Figure 3. **Qualitative Comparisons on 4P Composition Benchmark.** Under the *Spatial Canvas* setup, our Canvas-to-Image achieves the highest identity preservation for multi-subject insertion while respecting the spatial placement of each subject segment. FLUX Kontext [3]-based approach [18] fails to preserve identity, whereas NanoBanana [42] consistently exhibits copy-pasting artifacts. Compared to our base model, Qwen-Image-Edit [45], our method maintains similar image quality but demonstrates significantly stronger identity preservation.

tic embeddings, while the canvas image is also simultaneously encoded by the VAE into latents. These VLM embeddings, VAE latents, and noisy latents are concatenated and fed into the diffusion, which predicts the velocity for denoising. During training, we fine-tune the attention, image modulation, and text modulation layers in each block using LoRA [17] with a rank of 128. Note the feed-forward layers are frozen, as we find it is important to preserve the prior image quality of the pretrained model. Optimization is performed with AdamW [24], using a learning rate of 5×10^{-5} and an effective batch size of 32. The model is trained for 200K steps on 32 NVIDIA A100 GPUs.

Dataset. Our training is constructed from two primary data sources. The *Spatial Canvas* and *Pose Canvas* variants are derived from a large-scale internal, human-centric dataset containing 6M cross-frame images from 1M unique identities. This dataset enables flexible composition sampling for our *Multi-Task Canvas* formulation, for example, pairing subjects and backgrounds drawn in a cross-frame manner to avoid copy-pasting artifacts. See *Appendix* for details. For the *Box Canvas*, we extend the internal data with bounding box annotations from the external CreatiDesign

dataset [51], which provides a large-scale corpus of images annotated with boxes and named entities. During training, we sample each task type and its dataset with an uniform distribution for a balanced multi-task supervision.

Benchmarks. We benchmark our method against several baselines, including the base model Qwen-Image-Edit [45], the state-of-the-art commercial editing model Gemini 2.5 Flash Image (also known as Nano-Banana) [42], and other most recent related work such as CreatiDesign [51] and Overlay Kontext [18] in corresponding benchmarks. For a fair and direct comparison of unified-interface methods, our main paper evaluates baselines that also operate on a single image input. We provide an extended comparison against other methods such as ID-Patch [53] in *Appendix*. Evaluations are conducted across four distinct benchmarks: *(i)* 4P Composition via the *Spatial Canvas*, *(ii)* Pose-Overlaid 4P Composition via the *Pose Canvas*, *(iii)* the *Layout-Guided Composition* benchmark via the *Box Canvas*, and *(iv)* our proposed *Multi-Control Benchmark*, which is curated from the CreatiDesign benchmark [51] containing humans in prompts and augmented with our Spatial and Pose Canvas for reference subject in-

Figure 4. **Qualitative Comparisons on Pose-Overlaid 4P Composition Benchmark.** Our Canvas-to-Image achieves the highest identity preservation and most accurate pose alignment. Note how Canvas-to-Image closely follows the target poses defined in the prior generated by FLUX-Dev [2] ("Pose Prior" column), while maintaining subject identities more faithfully than the baselines.

jection and pose controlling. See *Appendix* for more details.
Metrics. We report ArcFace ID Similarity [8] for identity preservation, HPSv3 [25] for image quality, VQAScore [23] for text-image alignment. In addition, to assess the fidelity w.r.t. applied control (e.g. identity, pose, box), we introduce a Control-QA score (evaluated by an LLM). For each control, Control-QA assesses each image between a score of 1-to-5, depending on how aligned each generation to the given set of control combinations. We provide details of the Control-QA in *Appendix*. A comprehensive user study further validating these results is also provided in *Appendix*.

4.2. Qualitative and Quantitative Results

We present qualitative comparisons across the four benchmark setups in Figures 3-6. In the *4P Composition* benchmark (Fig. 3), Canvas-to-Image demonstrates superior identity preservation and spatial alignment when composing multiple personalized subjects, outperforming state-of-the-art baselines including Qwen-Image-Edit [45], the commercial model Nano-Banana [42], and Overlay Kontext [18], which is trained upon FLUX Kontext [3]. Nano-Banana consistently produces copy-pasted human segments, an observation supported by the quantitative results in Tab. 1.

Table 1. **Quantitative Comparison** of our method against baselines across four different control tasks. We report ArcFace ID Similarity [8] for identity preservation, HPSv3 [25] for image quality, VQAScore [23] for text–image alignment, and Control-QA for control adherence. The best results for each task are highlighted in **bold**, where the second best is highlighted as underlined.

Method	ArcFace ↑	HPSv3 ↑	VQAScore ↑	Control-QA ↑
4P Composition				
Qwen-Image-Edit [45]	0.258	13.136	0.890	3.688
Nano Banana [42]	0.434	10.386	0.826	3.875
Overlay Kontext [18]	0.171	12.693	0.879	2.000
Ours	**0.592**	**13.230**	**0.901**	**4.000**
Pose Guided 4P Composition				
Qwen-Image-Edit [45]	0.153	**12.940**	0.890	4.031
Nano Banana [42]	0.262	9.973	0.861	3.438
Ours	**0.300**	12.899	**0.897**	**4.469**
Layout-Guided Composition				
Qwen-Image-Edit [45]	-	10.852	0.924	3.813
Nano Banana [42]	-	10.269	0.917	3.750
CreatiDesign [51]	-	9.790	0.923	**4.844**
Ours	-	**10.874**	**0.935**	**4.844**
Multi-Control Composition				
Qwen-Image-Edit [45]	0.204	**12.251**	0.903	3.575
Nano Banana [42]	0.356	11.370	0.873	3.625
Ours	**0.375**	12.044	**0.906**	**4.281**

Such artifacts may occur because closed-source models such as Nano-Banana [42] are likely not trained with
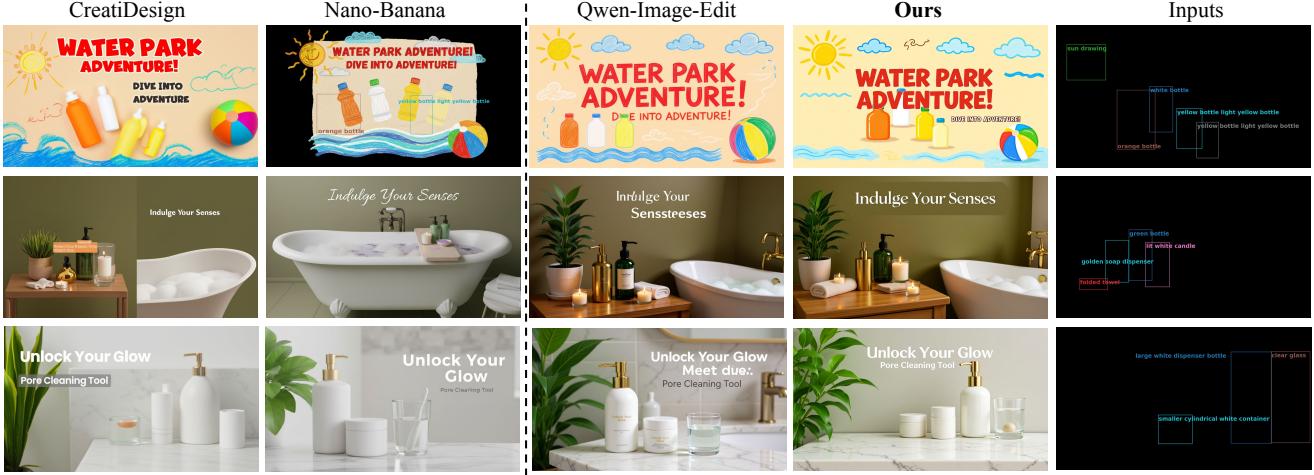
Figure 5. **Qualitative Comparisons on the Layout-Guided Composition Benchmark.** Under the *Box Canvas* setup, our Canvas-to-Image achieves the highest fidelity in spatial layout control, even compared to the state-of-the-art CreatiDesign [51] model trained for this task. Nano Banana [42], while demonstrating good image quality, does not adhere to the bounding boxes as closely as our model. Compared to our base model Qwen-Image-Edit [45], we achieve the same level of image quality but significantly stronger spatial condition alignment.
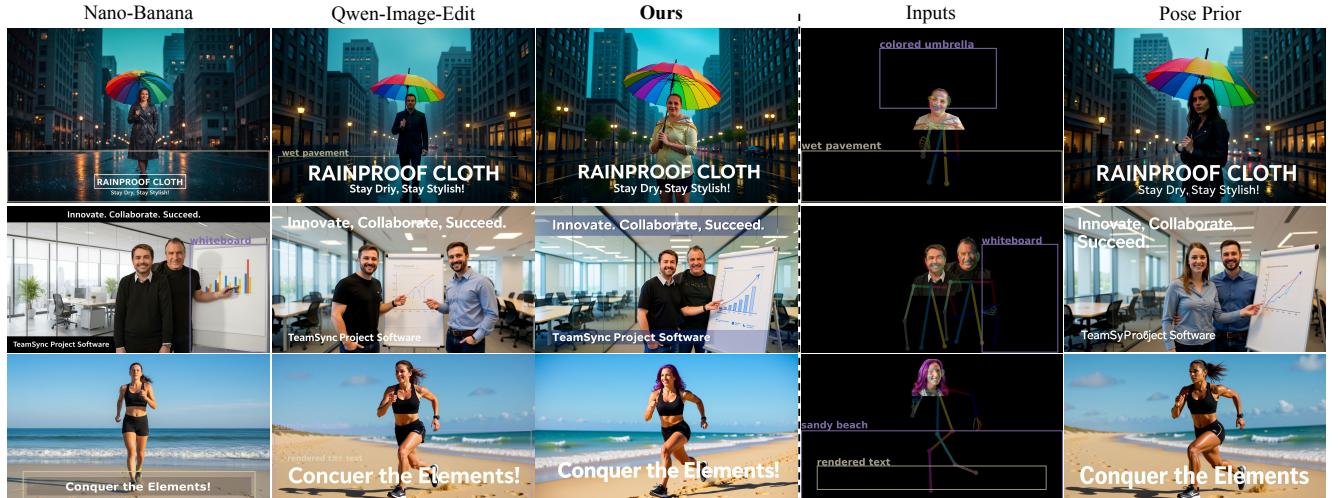

Figure 6. **Qualitative Comparisons on the Multi-Control Composition Benchmark.** We compare Canvas-to-Image with state-of-the-art baselines under inputs containing multiple heterogeneous control signals. Existing methods [42, 45] fail to simultaneously satisfy all conditions, often neglecting spatial, pose, or identity constraints. In contrast, Canvas-to-Image accurately adheres to the bounding boxes for spatial placement, respects pose and interaction cues from overlaid skeletons, and maintains strong identity fidelity of the reference identity images across multi-control inputs.

segment-like inputs, which are explicitly incorporated in our canvas-based training. Overlay Kontext and Qwen-Image-Edit [45] also fail to preserve subject identities (*e.g.*, 1st row 4th ID, 2nd row 3rd ID, and 3rd row 4th ID), a weakness reflected in their low ArcFace scores in Tab. 1.

In the benchmark with extra overlaid poses (Fig. 4), Canvas-to-Image is the only method that accurately follows the target poses ("Pose Prior" column) while maintaining high identity fidelity and visual realism, substantially outperforming baselines [42, 45]. For the *Layout-Guided Composition* benchmark (Fig. 5), Canvas-to-Image produces semantically coherent compositions that adhere to the box constraints, whereas Nano-Banana and Qwen-Image-Edit often ignore structural signals or suffer from annotation rendering artifacts. Notably, Canvas-to-Image also surpasses the dedicated state-of-the-art model Cre-
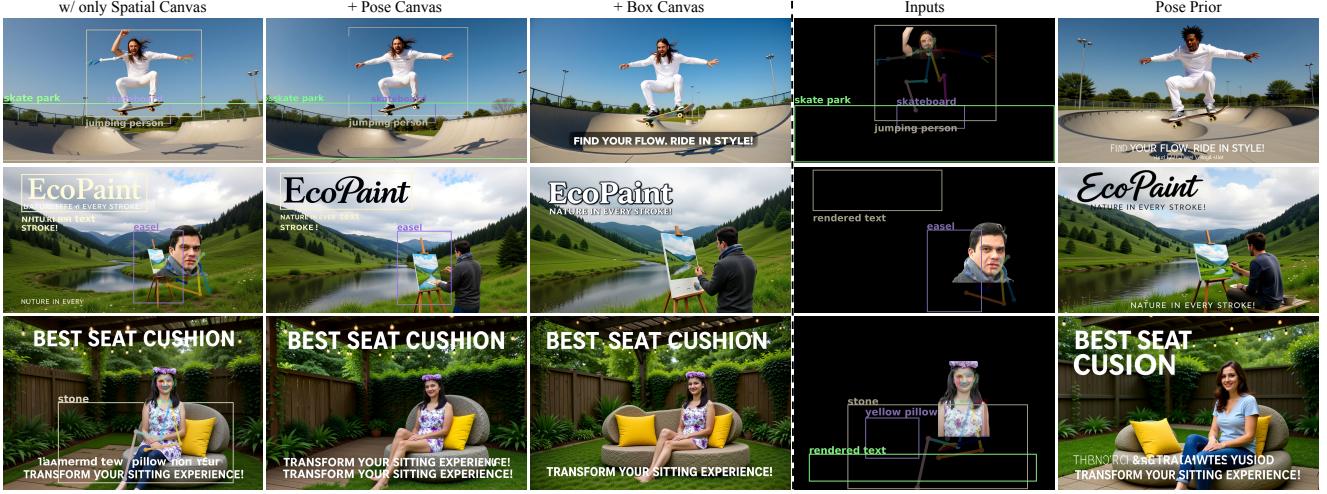
Figure 7. **Qualitative Ablation for Multi-Task Canvas Training.** Starting from training with only the Spatial Canvas, the model struggles to follow pose and bounding-box annotations. As we incrementally add the Pose Canvas and Box Canvas tasks, the model progressively learns to respect these additional controls. The final model effectively handles complex multi-control inputs. Notably, during training, each sample contains only a single control type, yet the model exhibits strong generalization to multi-control scenarios at inference.

Table 2. **Ablation study for Multi-Task Canvas Training.** Performance on the Multi-Control Benchmark is evaluated as the *Pose Canvas* and *Box Canvas* tasks are incrementally added to the baseline *Spatial Canvas* model.

Model	ArcFace↑	VQAScore↑	HPSv3↑	Control-QA↑
Spatial Canvas	**0.389**	0.865	10.786	4.156
+ Pose Canvas	0.371	0.874	11.440	4.188
+ Box Canvas	0.375	**0.906**	**12.044**	**4.281**

atiDesign [51], which was trained specifically for this task in the training set of CreatiDesign evaluation benchmark.

Finally, on the *Multi-Control Benchmark* (see Fig. 6), where identity preservation, pose guidance, and box annotations must be satisfied jointly, our model achieves the highest compositional fidelity. It integrates reference subjects and multiple control cues seamlessly, while baselines [42, 45] often produce artifacts or fail to satisfy all input constraints. Quantitatively, Tab. 1 validates the effectiveness of our unified framework. The balanced performance across control adherence and identity preservation confirms that encoding heterogeneous signals into a single canvas successfully enables the simultaneous execution of spatial, pose, and identity constraints.

We highlight that all results across benchmarks are generated by the same unified Canvas-to-Image model, demonstrating its strong generalization from single-control training samples to complex control scenarios at inference.

4.3. Ablation Studies

We conduct ablation studies to evaluate the effectiveness of our **Multi-Task Canvas Training** on the *Multi-Control Benchmark*. We start with a baseline model trained *only* on the *Spatial Canvas* and then progressively add the *Pose Canvas* and *Box Canvas* tasks to the training curriculum. Quantitative and qualitative results are presented in Tab. 2 and Fig. 7, respectively. Tab. 2 clearly show that as more canvas tasks are incorporated, we observe consistent gains in image quality (HPSv3) and control adherence (Control-QA). The qualitative results (Fig. 7) confirm this: the baseline model fails to follow pose and layout instructions, while the full model successfully handles all multi-control inputs. Additionally, we provide ablations on the impact of the trained branches of MM-DiT[31] and the convergence behavior of control following in *Appendix*.

5. Conclusion

We introduced Canvas-to-Image, a unified framework for flexible, compositional image generation. Our approach enables a diffusion model to reason jointly over reference subjects, pose signals, and layout constraints by reformulating these heterogeneous controls into a single canvas-conditioned paradigm. Our Multi-Task Canvas training enables Canvas-to-Image to generalize from single-control training samples to complex multi-control scenarios at inference, allowing a single unified model to achieve strong identity preservation, pose fidelity, and structural coherence. This unified canvas formulation establishes a scalable paradigm for multi-modal guidance; while currently bounded by the information density of a single RGB interface, as discussed in the *Appendix*, it establishes a robust foundation for future work to enable even richer forms of visual and semantic control.
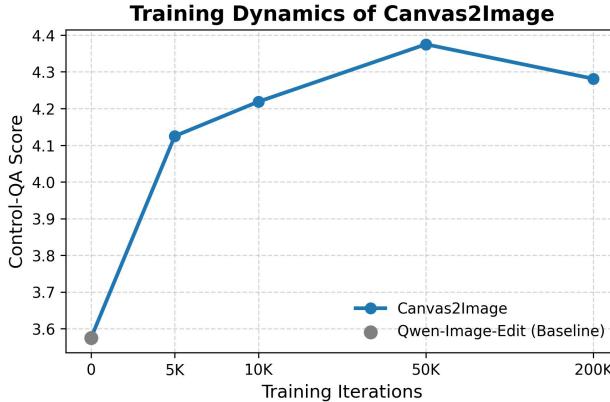
8

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 3

[2] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 6, 5, 7, 8

[3] Black Forest Labs. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 5, 6, 1

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 2, 4

[5] 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. Multi-subject open-set personalization in video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 6099–6110. Computer Vision Foundation / IEEE, 2025. 3

[6] Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. Lorashop: Training-free multi-concept image generation and editing with rectified flow transformers. *CoRR*, abs/2505.23758, 2025. 3

[7] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging properties in unified multimodal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[8] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. 6, 1, 7

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*. OpenReview.net, 2024. 2

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023. 2

[11] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions on Graphics (TOG)*, 44(4):1–11, 2025. 3

[12] 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. Preventing shortcuts in adapter training via providing the shortcuts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2

[13] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning ID customization via contrastive alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[14] Yucheng Han, Rui Wang, Chi Zhang, Juntao Hu, Pei Cheng, Bin Fu, and Hanwang Zhang. EMMA: your text-to-image diffusion model can secretly accept multi-modal prompts. *CoRR*, abs/2406.09162, 2024. 2, 3

[15] Junjie He, Yifeng Geng, and Liefeng Bo. 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)*, 2025. 3, 1

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2022. 5

[18] ilkerzgi and gokaygokay. Overlay-kontext-dev-lora. https://huggingface.co/ilkerzgi/Overlay-Kontext-Dev-LoRA, 2025. LoRA fine-tune of FLUX.1-Kontext-dev for image overlay tasks. 5, 6, 1, 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 2, 7

[20] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 253–270. Springer, 2024. 3

[21] 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 (CVPR)*, pages 1931–1941, 2023. 2

[22] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023. 2, 3

[23] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan.

Evaluating text-to-visual generation with image-to-text generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–384. Springer, 2024. 6, 1, 7

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[25] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 6, 1, 7

[26] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 2, 3

[27] 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. Dreamo: A unified framework for image customization. In *SIGGRAPH Asia 2025 Conference Papers*, 2025. 1, 2

[28] OpenAI. Gpt-4o technical report, 2024. Accessed: 2025-05-22. 7

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. 1, 2, 7

[30] Or Patashnik, Rinon Gal, Daniil Ostashev, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. 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*, pages 1–12, 2025. 2

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 8

[32] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7964–7973. IEEE, 2024. 3

[33] 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. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. 2

[34] Guocheng Qian, Kuan-Chieh Wang, Or Patashnik, Negin Heravi, Daniil Ostashev, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Omni-id: Holistic identity representation designed for generative tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[35] Guocheng Gordon Qian, Daniil Ostashev, Egor Nemchinov, Avihay Assouline, Sergey Tulyakov, Kuan-Chieh Jackson Wang, and Kfir Aberman. Composeme: Attribute-specific image prompts for controllable human image generation. In *SIGGRAPH Asia 2025 Conference Papers*, 2025. 2, 3

[36] Guocheng Gordon Qian, Ruihang Zhang, Tsai-Shien Chen, Yusuf Dalva, Anujraaj Goyal, Willi Menapace, Ivan Skorokhodov, Daniil Ostashev, Meng Dong, Arpit Sahni, Ju Hu, Sergey Tulyakov, and Kuan-Chieh Jackson Wang. Layercomposer: Multi-human personalized generation via layered canvas. *arXiv*, 2025. 3

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[38] 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 (CVPR)*, pages 10684–10695, 2022. 2

[39] 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 (CVPR)*, pages 22500–22510, 2023. 2

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[42] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. 5, 6, 7, 8, 1, 2

[43] Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[44] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2

[45] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *CoRR*, abs/2508.02324, 2025. 2, 4, 5, 6, 7, 8, 1

[46] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *CoRR*, abs/2506.18871, 2025. 1, 2

[47] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1, 2

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

[49] XLabs-AI. Flux-controlnet collections. `https://huggingface.co/XLabs-AI/flux-controlnet-collections`, 2024. Accessed: 2024-11-13. 2

[50] 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. 2

[51] Hui Zhang, Dexiang Hong, Maoke Yang, Yutao Cheng, Zhao Zhang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatidesign: A unified multi-conditional diffusion transformer for creative graphic design. *arXiv preprint arXiv:2505.19114*, 2025. 2, 3, 5, 6, 7, 8, 1

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3

[53] Yimeng Zhang, Tiancheng Zhi, Jing Liu, Shen Sang, Liming Jiang, Qing Yan, Sijia Liu, and Linjie Luo. Id-patch: Robust ID association for group photo personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2986–2996. Computer Vision Foundation / IEEE, 2025. 2, 3, 5, 1

[54] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 3

[55] Zhengguang Zhou, Jing Li, Huaxia Li, Nemo Chen, and Xu Tang. Storymaker: Towards holistic consistent characters in text-to-image generation. *arXiv preprint arXiv:2409.12576*, 2024. 2, 3

Canvas-to-Image: Compositional Image Generation with Multimodal Controls

Supplementary Material

Figure I. **Training Dynamics for Canvas-to-Image.** The Control-QA score steadily improves during early training and converges around 50K iterations, indicating that the model effectively learns consistent control and composition. We further train up to 200K iterations to refine local details and enhance robustness in generation quality.

A. Comparisons with Personalization Methods

In the main paper, we compare Canvas-to-Image primarily with approaches designed to naturally support the composition task with free form inputs, such as [18, 42, 45, 51]. However, acknowledging that composition and personalization are intersecting tasks, we provide supplementary comparisons with recent zero-shot personalization methods capable of supporting multiple-concept input, specifically UniPortrait [15], FLUX Kontext [3], UNO [47], OmniGen2 [46], DreamO [27], and ID-Patch [53]. To ensure a robust evaluation, we extend these comparisons across three distinct benchmarks, detailed below.

4P Composition Benchmark. Following the 4P Composition setup introduced in the main paper, we provide quantitative comparisons with personalization methods, including [15, 27, 46] and [3] in Tab. I. We validate that Canvas-to-Image achieves superior identity preservation (ArcFace), image quality (HPSv3), and text-image alignment (VQAScore) compared to these competing baselines. In addition to the quantitative metrics, we provide comparative qualitative examples in Fig. II for a visual assessment. These examples illustrate how Canvas-to-Image achieves more consistent identity preservation and realistic multi-subject composition compared to all personalization-based baselines as well as Qwen-Image-Edit [45].

Table I. **Quantitative Comparison Including Personalization Baselines on the 4P Composition Benchmark. Bold** values denote the best performance for each metric, highlighting the superior overall performance of our method across all categories.

	ArcFace ↑	HPSv3 ↑	VQAScore ↑	Control-QA ↑
DreamO [27]	0.2049	12.4210	0.7782	1.4062
OmniGen2 [46]	0.0859	12.9873	0.8051	1.9688
ID-Patch [53]	0.0824	7.1262	0.7846	1.0938
UniPortrait [15]	0.3088	12.4011	0.7860	2.5000
Overlay Kontext [18]	0.1709	12.6932	0.8792	2.0000
Flux Kontext [3]	0.2168	12.7684	0.8687	2.2188
UNO [47]	0.0769	12.1558	0.8402	1.5000
Nano Banana [42]	0.4335	10.3857	0.8260	3.8750
Qwen Image Edit [45]	0.2580	13.1355	0.8974	3.6875
Ours	**0.5915**	**13.2295**	**0.9002**	**4.0000**

Table II. **Quantitative Comparisons on the Pose-Guided 4P Composition Benchmark.** The Control-QA score provides a unified criterion that accounts for both pose accuracy and identity preservation, where our method achieves the highest performance among all baselines.

Pose	ArcFace ↑	HPSv3 ↑	VQAScore ↑	Control-QA ↑	PoseAP$_{0.5}$ ↑
ID-Patch [53]	<u>0.2854</u>	11.9714	<u>0.8955</u>	<u>4.1250</u>	**75.0814**
Nano Banana [42]	0.2623	9.9727	0.8609	3.4375	64.1704
Qwen-Image-Edit [45]	0.1534	**12.9397**	0.8897	4.0312	67.2734
Ours	**0.3001**	12.8989	**0.8971**	**4.4688**	<u>70.1670</u>

Table III. **Quantitative Results on the ID-Object Composition Benchmark.** We compare our method with several baselines across five different metrics. **Bold** values indicate the best performance in each column. DINOv2 measures object preservation. Our Canvas-to-Image achieves the highest identity (ArcFace) and object (DINOv2 [29]) preservation as well as the highest control following (Control-QA).

	ArcFace ↑	HPSv3 ↑	VQAScore ↑	Control-QA ↑	DINOv2 ↑
UNO [47]	0.0718	8.6718	0.8712	2.5000	0.2164
DreamO [27]	0.4028	9.0394	0.8714	3.9688	0.3111
OmniGen2 [46]	0.1004	10.2854	**0.9266**	4.4062	0.3099
Overlay Kontext [18]	0.1024	8.6132	0.8539	3.2812	0.2703
Flux Kontext [3]	0.1805	9.2179	0.8914	3.1562	0.2818
Qwen-Image-Edit [45]	0.3454	**10.3703**	0.9045	4.4062	0.2867
Ours	**0.5506**	9.7868	0.9137	**4.8750**	**0.3298**

Pose-Guided 4P Composition Benchmark. We provide additional quantitative and qualitative comparisons with ID-Patch [53], a method specifically designed for pose-guided composition with human identities. These results are detailed in Table II.

To rigorously evaluate this task, we employ a comprehensive set of metrics: ArcFace [8] for identity, HPSv3 [25] for aesthetic quality, VQAScore [23] for semantic alignment, and our proposed Control-QA score (see Sec. F). Furthermore, we introduce the PoseAP$_{0.5}$ score, which reports the Average Precision (AP$_{0.5}$) for extracted pose keypoints

1

Figure II. **Supplementary Qualitative Comparisons on the 4P Composition Benchmark with Personalization Approaches.** Both Qwen-Image-Edit and our Canvas-to-Image significantly outperform the state-of-the-art multi-subject personalization baselines DreamO [27], OmniGen2 [46], and UNO [47] in terms of identity preservation. Compared to Qwen-Image-Edit, our method demonstrates further improvements in identity fidelity, particularly for the rightmost man in the 1st row, the leftmost woman in the 2nd row, and the second man in the 3rd row.

to strictly measure spatial adherence.

As shown in Table II, Control-QA provides a unified evaluation criterion that jointly considers pose accuracy and identity preservation (measured by ArcFace similarity), under which our method achieves the highest score. Qualitative results in Fig. III further demonstrate that although ID-Patch [53], through its integration with Control-Net [52], can effectively reproduce target poses—resulting in high PoseAP scores—it often fails to maintain the correct number of subjects and consistent identities. In contrast, Canvas-to-Image achieves a more balanced trade-off between pose fidelity and identity preservation. Additional qualitative examples of pose-guided composition in single-person (1P) and two-person (2P) scenarios are presented in Fig. VI and Fig. V, respectively.

ID-Object Interaction Benchmark. To demonstrate the generalizability of our approach beyond human subjects, and to evaluate performance in scenarios involving natural interactions between subjects, we extend our evaluations to the ID-Object Interaction benchmark. To construct this benchmark, we pair human identities from the FFHQ-in-the-Wild [19] dataset with object references from the DreamBooth [39] dataset to create challenging ID–Object pairs.

We quantitatively compare our method against a wide range of baselines, including [3, 18, 27, 42, 46, 47], as well as our main baseline, Qwen-Image-Edit [45]. Corresponding quantitative results are provided in Table III. Our Canvas-to-Image achieves the highest identity and object preservation, as well as the strongest overall control following, as indicated by ArcFace, DINOv2 [29], and Control-QA metrics, respectively. To further assess in-

Figure III. **Supplementary Qualitative Comparisons on the Pose-Overlaid 4P Composition Benchmark.** We additionally include the relevant state-of-the-art personalization baseline, ID-Patch [53], in our comparison. While ID-Patch follows poses to some extent, it performs significantly worse in identity preservation and image quality. In contrast, image editing baselines fail to accurately follow the target pose. Our Canvas-to-Image achieves both strong identity preservation and precise pose alignment.

teraction fidelity, we provide qualitative comparisons in Fig. IV. Canvas-to-Image produces coherent compositions that faithfully preserve both human identity and object fidelity, maintaining correct proportions and natural interactions between them, whereas existing baselines often fail to achieve realistic integration of the two.

B. Supplementary Ablations

In addition to the ablation studies in the main paper, we provide a deeper analysis of the training dynamics of Canvas-to-Image. Specifically, we examine the convergence behavior under multi-task learning and empirically validate our selection of trainable blocks.

Convergence Behavior of Canvas-to-Image. We tracked the model's performance across different training iterations (Fig. I). The Control-QA curve shows steady improvement in the early stages, with rapid gains up to

50K iterations, where convergence is largely achieved. During this phase, the model progressively strengthens control adherence. Although key metrics plateau beyond 50K, we continue training up to 200K iterations to refine local details and improve robustness. All subsequent ablation studies use this 200K-iteration model as the default checkpoint.

Ablations of Trainable Blocks. We investigate the impact of different architectural choices for LoRA optimization. In our default configuration, we train modulation and attention layers within the text and image attention branches, while keeping feed-forward layers frozen. Table IV quantifies the impact of including or excluding these components on the 4P Composition benchmark. Two key findings emerge from this analysis. First, effective identity preservation requires the joint training of both the text and image branches; omitting either leads to a drop in identity

Figure IV. **Qualitative Results on the ID–Object Composition Benchmark.** Our Canvas-to-Image generates coherent compositions that faithfully preserve both human identity and object fidelity, maintaining correct proportions and natural interactions between them. In contrast, existing baselines often fail to achieve realistic integration between the human and the object. For instance, the baseline Qwen-Image-Edit [45] fails to preserve both identity and object consistency, as illustrated in these examples.

fidelity. Second, training the feed-forward layers negatively impacts the model's generalization; we observe a deterioration in both visual quality and prompt alignment when these layers are unfrozen. Based on these results, our final model excludes feed-forward layers from the optimization process. Finally, we evaluate the contribution of the task indicator prompt (c). As reported in Tab. IV, removing this indicator leads to a degradation in performance across all metrics. This confirms that explicitly signaling the control type is crucial for the model to resolve ambiguity and effectively switch between different compositional reasoning modes. We provide qualitative ablations on the task indicator in Fig. VII
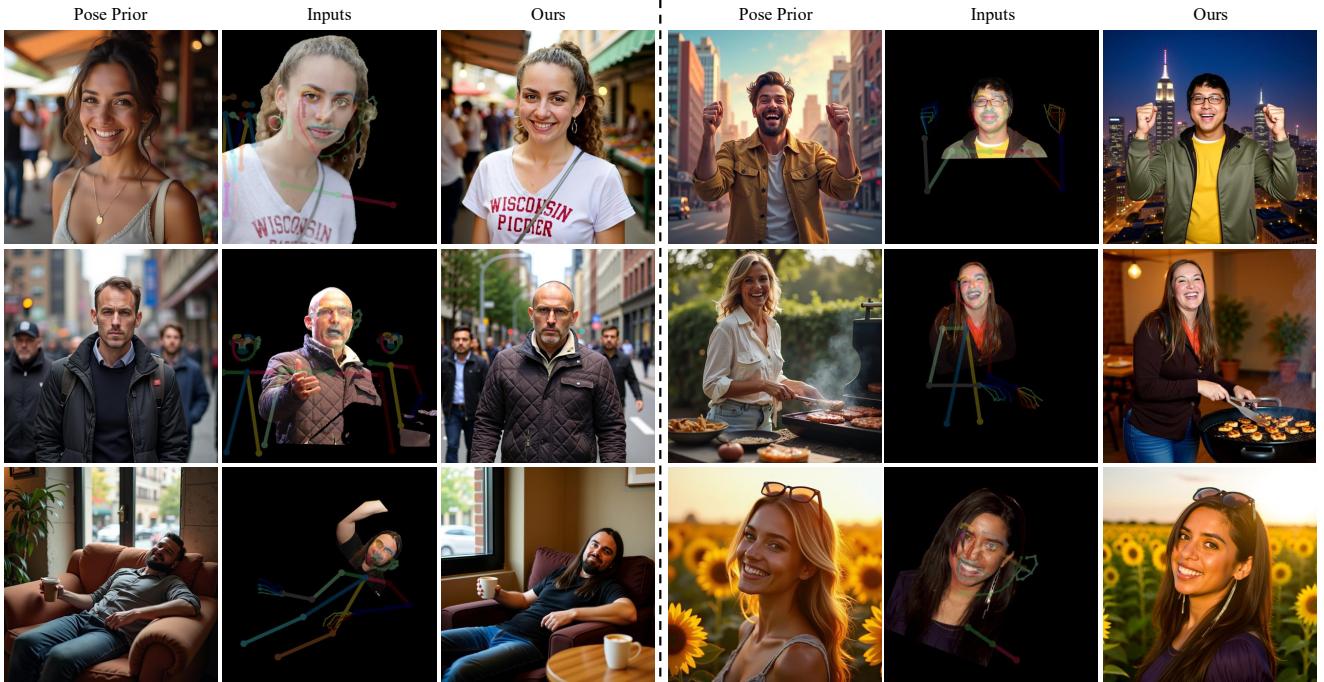
C. Limitations

While Canvas-to-Image provides an intuitive interface for compositional image generation with multimodal inputs, enabling combined controls in a single inference pass, the "visual canvas" format has inherent constraints. Although this format offers significant advantages in usability and flexibility, it is strictly bound by the available pixel space. As demonstrated in Fig. 3 and 4, Canvas-to-Image successfully handles occluding entities up to 4P composition,

Figure V. **Qualitative Results on Pose-Overlaid 2P Composition.** Under the *Pose Canvas* setup, our Canvas-to-Image achieves superior identity preservation and accurate pose alignment. Notably, Canvas-to-Image closely follows the target poses defined by the prior generated from FLUX-Dev [2] ("Pose Prior" column), while producing coherent and high-quality images.

Figure VI. **Qualitative Results on Pose-Overlaid 1P Composition.** Under the *Pose Canvas* setup, our Canvas-to-Image again achieves superior identity preservation and accurate pose alignment.

outperforming baseline approaches. However, relying on a single RGB canvas implicitly limits the number of concepts

Table IV. **Ablations on Model Architecture.** We conduct ablations on the fine-tuned layers using the 4P Composition Benchmark. Our default configuration, which fine-tunes both the text and image branches while excluding the feed-forward layer, with task indicators included, achieves the highest overall performance.

Model	ArcFace↑	HPSv3↑	VQAScore↑
Qwen-Image-Edit	0.2580	13.1355	0.8974
Ours w/o Text Branch	0.4917	11.6523	0.8297
Ours w/o Image Branch	0.4687	12.7077	0.8880
Ours w/ Feed-Forward	0.5603	12.4846	0.8577
Ours w/o Task Indicator	0.5217	12.6046	0.8555
Ours	**0.5915**	**13.2295**	**0.9002**

Figure VII. **Qualitative Ablations on the Task Indicator.** We visualize the impact of removing the task indicator prompt (c) in training. Without this explicit signal, the model suffers from task mix-up, where the 4P Composition (Spatial Canvas) is impacted by the Box Canvas task. This results in unwanted text artifacts appearing in the background, as the model incorrectly transfers the text-rendering behavior required only in box-canvas settings to a spatial composition benchmark that does not require text rendering.

that can be interpreted simultaneously; as the number of concepts increases, the canvas becomes crowded and harder to interpret. To resolve this, future work could explore layered controls, such as designing the input canvas with an additional alpha channel (RGBA).

D. Additional Applications

Canvas-to-Image is also capable of background-aware composition. We provide qualitative examples of this capability in Fig. VIII. Canvas-to-Image can inject humans or objects into a scene through reference image pasting or bounding box annotation, with the inserted elements naturally interacting with the background.

E. User Study

We validate the effectiveness of Canvas-to-Image on the Multi-Control Composition task through human evaluation. To ensure a fair and accurate assessment, we conduct two separate user studies aimed at evaluating the condition-following behavior of Canvas-to-Image against competing methods. Given the cognitive difficulty of assessing three simultaneous conditions (pose, identity, and box layout) at once, we decouple the input controls into two distinct pairwise comparisons: Pose + Identity" and Pose + Box Layout".

For each combination, we perform a separate study with unique participants. In total, we collected responses from 30 anonymous participants for 30 examples per study, conducted via the *Prolific* platform. The specific setups are detailed below:

- **Control Following (Pose + Box Layout"):** This study focuses on the structural capabilities of the model. For each question, users are shown an input pose reference and a box layout. Then they are presented with generated samples and asked to select which generation better adheres to the input controls. We utilize an A/B testing setup in which users select their preferred output. The instructions provided to the participants are shown in Fig. IX, and a sample question is provided in Fig. X.
- **Identity Preservation (Pose + Identity"):** To evaluate identity fidelity under spatial constraints, this study focuses on how well the subject's identity is preserved while applying a specific pose. Users are instructed to prioritize identity preservation in their assessment while verifying that the pose is applied. Similar to the previous study, we use an A/B setup. User instructions are provided in Fig. XI, with a sample question in Fig. XII.
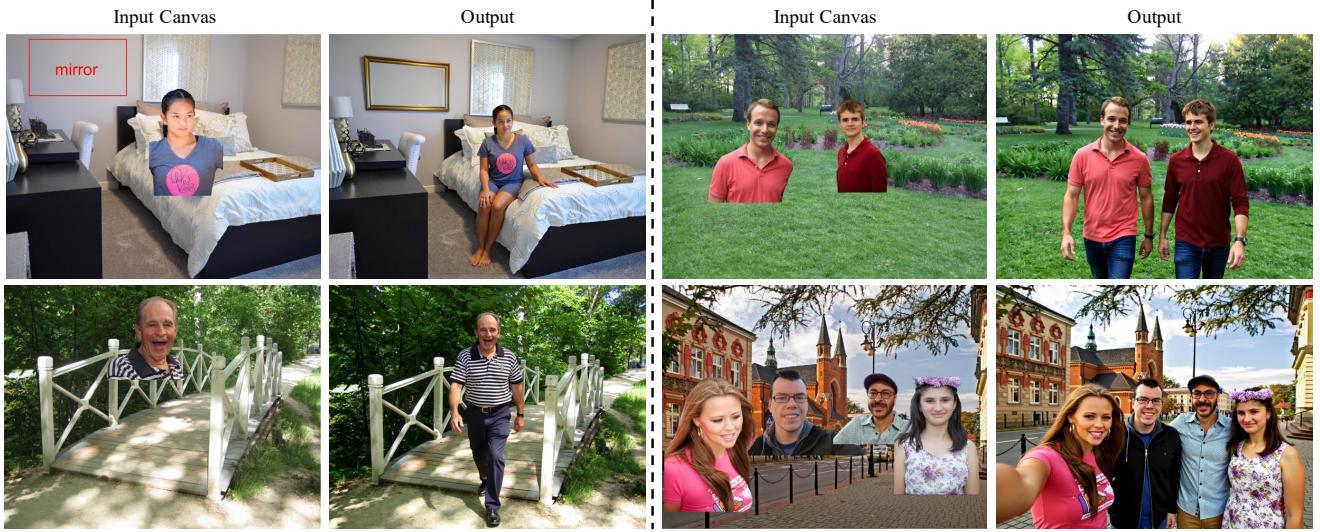
We report the win rates against competing methods in Table V for both the Control Following" and Identity Preservation" evaluations. Consistent with the quantitative analyses in the Multi-Control Composition Benchmark, we compare our method against [45] and [42].

Our results indicate that while Canvas-to-Image outperforms both baselines overall, there is a distinct trade-off among competitors: [42] performs stronger on identity preservation, whereas [45] performs better on control following. This alignment between human preference and our reported metrics serves as a strong validation for our proposed Control-QA score, confirming its success as a unified metric for evaluating multiple control inputs.

F. Benchmark Details

F.1. Evaluation Metrics

For all evaluations, we employ a unified setup focusing on identity preservation, visual quality, prompt alignment, and control adherence. We detail the specific metrics below.

Figure VIII. **Background-Aware Composition with Canvas-to-Image.** Given a background image, Canvas-to-Image seamlessly integrates humans or objects into the scene through reference image pasting or bounding box annotations, producing natural spatial alignment and consistent lighting with the surrounding environment.

Table V. **User Study Results.** The win rate represents the percentage of cases where users preferred our results over baseline methods. Canvas-to-Image is significantly preferred in both control following and identity preservation compared to strong baselines.

	Control Following	Identity Preservation
Ours vs. Qwen-Image-Edit [45]	**67.3%**	**77.3%**
Ours vs. Nano Banana [42]	**78.9%**	73.8%

ArcFace & DINOv2. We use ArcFace [8] to quantitatively evaluate identity preservation across all benchmarks involving human subjects (i.e., 4P Composition, Pose-Guided 4P Composition, Multi-Control Composition, and ID-Object Interaction). We sample input identities from the FFHQ-in-the-wild [19] dataset and compute the Arc-Face similarity score between the generated image and the masked identities in the corresponding composition. For object consistency in the ID-Object Interaction Benchmark, we utilize DINOv2 [29] in a similar manner to calculate similarity scores.

HPSv3 and VQAScore. To evaluate visual quality and adherence to the input prompt, we use the Human Preference Score v2 (HPSv3) [25]. We compute this metric using the original generation prompt. Although closed-source models like [42] may employ internal prompt rewriting, we ignore such implicit augmentations to ensure a fair comparison based on the user-provided input. Additionally, we utilize VQAScore [23] to further assess prompt alignment across our experiments.

Control-QA. Given the variety of control settings in Canvas-to-Image (e.g., pose, spatial layout, layout boxes), we establish a unified evaluation framework using an LLM-based scoring system. We employ GPT-4o [28] as a multimodal expert to rate the generated compositions against the provided control images. The system prompts used for the 4P Composition, Pose-Guided 4P Composition, Layout-Guided Composition, and Multi-Control Composition benchmarks are provided in Tables VI, VII, VIII, and IX, respectively. Note that additional quantitative evaluations for pose adherence (PoseAP) are discussed in Sec. A, and human user studies are detailed in Sec. E.

F.2. Evaluation Benchmarks

We employ an automated pipeline to construct the input canvases for all benchmarks. Below, we detail the construction process for each specific task.

4P Composition Benchmark. To construct the canvases for the 4P Composition benchmark, we randomly sample four human identities from the FFHQ-in-the-wild dataset [19]. To determine a natural spatial arrangement for these individuals, we employ a two-step process. First, we generate a synthetic "prior image" using FLUX.1-Dev [2] based on the target prompt. Second, we detect the human instances within this prior image to obtain realistic bounding boxes. Finally, we construct the input canvas by placing the segmented FFHQ identities into these extracted positions.

7

Pose-Guided 4P Composition Benchmark. Building upon the 4P Composition setup, we incorporate structural control into the pipeline. We utilize the same FLUX.1-Dev [2] prior images generated for the 4P task, but instead of just extracting bounding boxes, we utilize our internal pose estimation model to extract the target poses. We then construct the input canvas by placing these target poses alongside the reference identities.

Layout-Guided Composition Benchmark. As this benchmark focuses on named entity composition based on a layout rather than human identity, we utilize the test set of the CreatiDesign [51] dataset. Since our canvas format utilizes text overlaid directly on the image (rather than regional prompting), we filter the test set to select samples compatible with this modality. It is worth noting that the CreatiDesign dataset places a strong emphasis on text rendering capabilities, as demonstrated in our qualitative comparisons (see Fig. 5).

Multi-Control Composition Benchmark. For this complex setting, we leverage the text prompts and named entity annotations from the CreatiDesign [51] test set, specifically filtering for samples that involve human subjects. To obtain a valid target pose that aligns with these prompts, we generate a synthetic prior image using our baseline model, Qwen-Image-Edit [45]. Crucially, we do not utilize the pixel data of this prior image as a direct input; instead, we use it strictly to extract the target skeletal pose. We then construct the final input canvas by combining this extracted pose, a sampled reference identity, and the named entity annotations (for text rendering) from the original CreatiDesign sample. This setup simultaneously evaluates identity preservation, pose adherence, and text rendering, as highlighted in Fig. 6 and 7.

F.3. Dataset Details

To train Canvas-to-Image, we utilize an internal cross-frame dataset augmented with the CreatiDesign [51] dataset. Our internal dataset comprises ∼6M human-centric training images, constituting ∼1M scenes with cross-frame samples. Due to legal constraints, we cannot open-source this internal dataset; however, a similar multi-frame dataset can be constructed from public open-source video datasets. From these 1M scenes, we use an internal instance segmentation model to extract human segments for constructing the input canvases, while treating the remaining image areas as the background. Similarly, we extract poses from the target frames using an internal pose estimation model. Since Canvas-to-Image is built upon this human-centric data, we construct the human boxes in the "Box Canvas" using these extracted segments. To enable the model to be capable with a variety of objects, we include the CreatiDesign [51]

Figure IX. **User Instructions for User Study "Control Following".**

Figure X. **Sample Question for User Study "Control Following".**

Figure XI. **User Instructions for User Study "Identity Preservation".**

dataset, which introduces named annotations into our training set along with text-rendering focused samples.

Table VI. **Compositional Fidelity Evaluation Protocol (System Prompt).** This protocol was provided to the human evaluators and served as the instruction set for the LLM-based scoring system.

System Prompt Content

You are an expert visual analyst and quality assurance evaluator for an AI image generation system. Your task is to compare two images: an "Input Canvas" (Image 1) and a "Generated Scene" (Image 2).

Your goal is to provide a single, holistic score that judges the **Compositional Fidelity**. This score must be based on three *combined* criteria:

1. **Identity Preservation:** Do the individuals in the Generated Scene (Image 2) look like the correct people from the Input Canvas (Image 1)?
2. **Spatial Order:** Are these *same* individuals placed in the correct relative left-to-right order?
3. **Realism & Integration:** Do the individuals look like they *belong* in the scene? Or do they look "pasted on"? The lighting, shadows, and perspective on the subjects must be consistent with the new scene.

Your evaluation logic must link these three criteria. A scene with the right people in the right order, but looking like a bad "cut and paste" job, is a failure.

Tolerance: This is a composition, not a simple copy. Slight differences in pose, expression, or clothing are fine. Do not penalize minor artistic adjustments as long as the core **identity**, **relative order**, and **scene integration** are preserved.

Instructions:
1. Identify the individuals in the Input Canvas (Image 1) from left to right.
2. Find those same individuals in the Generated Scene (Image 2).
3. Evaluate how well the AI preserved all three criteria: Identity, Spatial Order, and Realism.
4. Provide a single, holistic score based on the rubric below.

Scoring Rubric (1-5):
* **5 (Excellent):** All individuals are clearly identifiable, they are in the correct relative order, AND they are all **flawlessly integrated** into the scene (correct lighting, shadows, scale, and **no cutout artifacts**).
* **4 (Good):** All three criteria are met, but with a minor flaw in *one* area (e.g., one identity is slightly weak, one person has slightly mismatched lighting, OR one minor spatial swap). Still free of obvious artifacts.
* **3 (Partial):** A significant flaw in one criterion OR minor flaws in several. For example, identities and order are correct, but the subjects look **pasted on** (poor realism, **faint but visible cutout edges**, or bad lighting). OR, realism is good, but identities/order are wrong.
* **2 (Poor):** Fails on at least two of the three criteria. OR, the scene **prominently displays cutout artifacts**, even if identity and order are correct.
* **1 (Failure):** The Generated Scene bears no meaningful resemblance to the Input Canvas, or is a clear "cut and paste" job with no integration.

Output Format:

Composition Fidelity Score: <A single numerical rating from 1-5>
Reasoning: <A brief explanation for your score. Justify your rating by referencing how well Identity, Spatial Order, AND Realism (including any cutout artifacts) were achieved *together*.>

Table VII. **Compositional Fidelity Evaluation Protocol (System Prompt) with Pose Control.** This protocol extends the previous evaluation by adding Pose Fidelity as a fourth critical criterion.

System Prompt Content

You are an expert visual analyst and quality assurance evaluator for an AI image generation system. Your task is to compare three images to judge the quality of a generated scene.

Your Inputs:
* **Image 1 (Pose Prior):** Shows the target pose skeletons (e.g., OpenPose).
* **Image 2 (Canvas):** Contains the subject cutouts. This defines **WHO** the person is (identity) and their **relative left-to-right order**.
* **Image 3 (Generated Scene):** The AI's final output.

Your Goal:
Provide a single, holistic score for **Compositional Fidelity**. This score must be based on **four** combined criteria:

1. **Identity Preservation** (from Image 2): Do the people in the Scene look like the people from the Canvas?
2. **Spatial Order** (from Image 2): Are the people in the correct left-to-right order?
3. **Pose Fidelity** (from Image 1): Are the people in the Scene matching the target poses?
4. **Realism & Integration**: Does the final image look natural? Or does it look like a "pasted on" collage with bad lighting or perspective?

Evaluation Logic (Very Important):
* All four criteria are linked. A failure in one is a failure for the composition.
* You must use the **left-to-right position** to link the images. The pose on the *left* in Image 1 applies to the person on the *left* in Image 2, and both should appear on the *left* in Image 3.
* A correct pose on the wrong person is a failure.
* The right person in the right pose but looking "pasted on" is a failure.
* The right person, right pose, right order, but "pasted" is a failure.

Tolerance: This is a composition. Slight, artistic differences in pose, expression, or clothing are fine. Do not penalize minor adjustments as long as the core **identity**, **order**, **pose intent**, and **realism** are preserved.

Scoring Rubric (1-5):

* **5 (Excellent):** All four criteria are met perfectly. Correct identities, correct order, correct poses, and realistic integration.
* **4 (Good):** A minor flaw in *one* of the four criteria (e.g., one pose is slightly off, one identity is weak, one person's lighting is bad, a minor order swap).
* **3 (Partial):** A major flaw in one criterion (e.g., all poses are wrong, or subjects look "pasted") OR minor flaws in several (e.g., weak identity *and* bad realism).
* **2 (Poor):** Fails on at least two criteria (e.g., wrong people *and* wrong poses, regardless of realism).
* **1 (Failure):** The Generated Scene bears no meaningful resemblance to the inputs.

Output Format:

Composition Fidelity Score: <A single numerical rating from 1-5>
Reasoning: <A brief explanation for your score. Justify your rating by referencing how well Identity, Order, Pose, AND Realism were achieved *together*.>

Table VIII. **Spatial Alignment Fidelity Evaluation Protocol (System Prompt).** This protocol focuses specifically on evaluating how well the model respects bounding box layouts and relative object positioning.

System Prompt Content

You are an expert visual analyst and quality assurance evaluator for an AI image generation system. Your task is to compare two images to judge the spatial alignment of specific elements.

Your Inputs:
   * **Image 1 (Spatial Layout):** This image shows bounding boxes with labels for specific objects (e.g., "circular window", "potted plant"). The *position* and *relative size* of these boxes define the expected layout.
   * **Image 2 (Generated Scene):** This is the AI's final output. It will contain a full scene, but should place the specified objects according to the Spatial Layout.

Your Goal:
Provide a single, holistic score for **Spatial Alignment Fidelity**. This score must solely reflect whether the objects identified in the "Spatial Layout" (Image 1) are present in the "Generated Scene" (Image 2) and appear in the **correct relative positions and proportions**.

Evaluation Logic:
   * **Focus ONLY on the boxed elements** and their relative positions, sizes, and orientations as suggested by the bounding boxes in Image 1.
   * **Ignore other elements** in Image 2 that are not specified in Image 1.
   * **Ignore artistic style, realism, or quality** of the generated objects themselves. The primary concern is whether the layout is matched.
   * The system understands that bounding boxes are approximations; minor deviations are acceptable for a high score, but significant shifts are not.
   * If an object specified in Image 1 is completely missing or unrecognizable in Image 2, that's a major penalty.

Scoring Rubric (1-5):

   * **5 (Excellent):** All specified objects are present and their relative positions, sizes, and general orientations perfectly match the Spatial Layout. The image is **free of any generation artifacts**, including the input bounding boxes.
   * **4 (Good):** All specified objects are present and mostly in the correct relative positions/sizes, with only one very minor deviation (e.g., one object is slightly shifted or scaled but clearly recognizable and in the right general area). Still free of artifacts.
   * **3 (Partial):** Most objects are present and correctly positioned, but one or two are significantly misplaced, incorrectly scaled, or one is missing. OR, the layout is correct but the scene **contains faint but visible traces** of the bounding boxes.
   * **2 (Poor):** Several objects are either missing, unrecognizable, or significantly misplaced. OR, the scene **prominently displays the bounding box artifacts**, even if the layout is partially correct.
   * **1 (Failure):** The Generated Scene bears no meaningful resemblance to the Spatial Layout in terms of the specified objects' placement.

Output Format:

Spatial Alignment Score: <A single numerical rating from 1-5>
Reasoning: <A brief explanation for your score, detailing which objects were correctly placed/sized and which were not. Mention if box artifacts were present.>

Table IX. **Joint Control Fidelity Evaluation Protocol (System Prompt).** This protocol evaluates the model's ability to handle three simultaneous control signals (Identity, Pose, and Spatial Layout) within a single input canvas.

System Prompt Content

```
    You are an expert visual analyst and quality assurance evaluator for an AI image generation system.  Your
task is to compare two images to judge how well a "Generated Scene" adheres to a "Combined Control Canvas".

    Your Inputs:
      * Image 1 (Combined Control Canvas):  This single image provides three types of control:
        1.  Identity:  The face of the person shown.
        2.  Pose:  The pose skeleton overlaid on the person.
        3.  Spatial Layout:  The labeled bounding boxes (e.g., "dress", "rendered text") showing where
elements should be.
      * Image 2 (Generated Scene):  This is the AI's final output.

    Your Goal:
    Provide a single, holistic score for Joint Control Fidelity.  This score must reflect how well the
Generated Scene simultaneously satisfies all control types.

    Evaluation Logic:
      * All criteria are linked.  A failure in one is a failure for the composition.
      * Identity:  Does the person in Image 2 look like the person in Image 1?
      * Pose:  Does the person's pose in Image 2 match the skeleton from Image 1?
      * Layout:  Are the elements from the bounding boxes (like "dress" or "rendered text") present in Image 2
in the correct locations?
      * Realism:  Does the final image look like a coherent, natural scene, or a "pasted" collage?

    A correct pose on the wrong person is a failure.  The right person in the right pose, but with text in the
wrong place, is a failure.  The right person, pose, and layout, but with a "pasted" look, is also a failure.

    ---

    Scoring Rubric (1-5):

      * 5 (Excellent):  All four criteria (Identity, Pose, Layout, Realism) are perfectly met.
      * 4 (Good):  A minor flaw in one of the four criteria (e.g., identity is slightly weak but recognizable,
text is a bit off-center, pose is almost right, minor lighting inconsistency).
      * 3 (Partial):  A major flaw in one criterion (e.g., pose is completely ignored, identity is wrong) OR
minor flaws in several.
      * 2 (Poor):  Fails on two or more criteria (e.g., wrong person and wrong pose).
      * 1 (Failure):  The Generated Scene bears no meaningful resemblance to the control inputs.

    ---

    Output Format:

    Joint Control Fidelity Score:  <A single numerical rating from 1-5>
    Reasoning:  <A brief explanation for your score, referencing Identity, Pose, Spatial Layout, and Realism.>
```
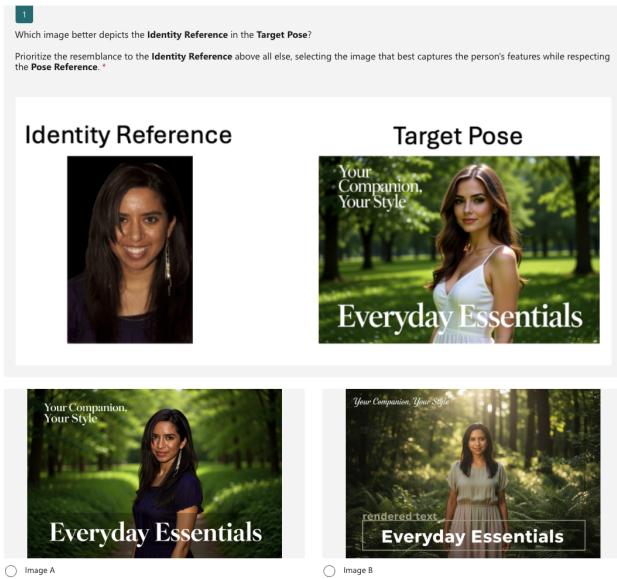
Figure XII. **Sample Question for User Study "Identity Preservation".**