CVPR
#9666

Disentangled Low-Rank Adaptation and Compositional Derivation for Controllable Spatial-Temporal Personalization

Anonymous CVPR submission

Paper ID 9666

(a) DisLoRA (b) Instance-Level Spatial Personalization (c) Event-Level Temporal Personalization

Figure 1. **Gallery of Spatial-Temporal Personalization with our D^{2}Plan**. We encapsulate scalable disentangled concepts in (a) DisLoRA set, and integrate two generative modules: (b) Instance-Level Spatial Personalization (ISP) controls local semantics (*e.g.*, "sunglasses" described by sub-prompt) while maintaining scene consistency (*i.e.*, layout determined by the global prompt); and (c) Event-Level Temporal Personalization (ETP) interpolates the consistent keyframes generated by ISP following the time-varying template decomposed by LLM.

Abstract

Recent personalization methods based on text-to-image diffusion models have demonstrated unprecedented synthesis quality, promoting the generation of personalized concepts even without pre-trained data. However, these concept-driven techniques fail to seamlessly extend concepts and control local fine-grained semantics in generating a long series of images, indicating the dilemma of representation entanglement. To track this issue, we propose the ***D^{2}Plan*** *that learns a semantic-**d**isentangled diffusion space, allowing scalable content **d**erivation for controllable spatial-temporal **p**erson**a**li**z**ati**on**. Specifically, we first introduce the Instance-Level Spatial Personalization (ISP) with the Disentangled Low-Rank Adaptation (DisLoRA) that encapsulates individual concepts within the disentangled space, ensuring non-degraded multi-concept scalability without fusion tuning. ISP also facilitates composition with fine-grained semantics while maintaining consistency, which follows the gradient-guided decomposition of the global and sub-prompts across multiple diffusion models. Moreover, we propose the Event-Level Temporal Personalization (ETP) that interpolates consistent keyframes acquired by ISP, accompanied by the time-varying template decomposed from the user-provided prompt using the Large Language Model (LLM). Additionally, our method is compatible with ControlNet, enabling more spatial conditions such as skeleton and depth. Extensive experiments demonstrate that our method outperforms recent state-of-the-art competitors in terms of identity preservation and controllability.*

1. Introduction

Recent diffusion models [10, 19] have made substantial advancements [1, 7, 23, 31, 33, 34] in the realm of text-to-image generation, benefiting from large-scale text-image paired datasets [4, 36]. Although these foundation models improve the quality and diversity of image generation with intuitive text prompts, they are unable to generate personalized concepts outside the dataset (*e.g.*, a specific man named "*John*"). Consequently, given several images of the target concept, personalization techniques [11, 35] are introduced to empower the foundational model in generating personalized concepts that align with the new context or style described by user-provided prompts.

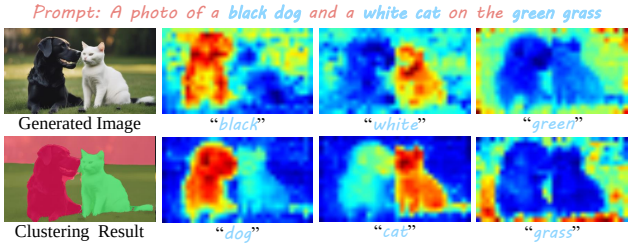


Figure 2. **Semantic-Disentangled Diffusion Space**. **Left lower**: Clustering results of cross-attention map. **Right three columns**: Average attention map across all time steps. The cross-attention maps observed from Stable Diffusion [34] exhibit strong text semantic and spatial alignment, indicating the model is capable of incorporating the disentangled words into the generation process.

Method	Multi-Concept	Plug-and-Play	Fine-Grained
DreamBooth [35]	✗	✓	✗
Textual Inversion [11]	✗	✓	✗
Custom Diffusion [22]	✓	✗	✗
Mix-of-Show [13]	✓	✗	✗
D^{2}Plan (Ours)	✓	✓	✓

Table 1. **Comparison of Previous Personalization Methods**. Our D^{2}Plan has three advantages: (1) *multi-concept* identity preservation with high quality; (2) *plug-and-play* multi-concept scalability; and (3) *local fine-grained control* under consistency.

Previous studies [11, 35] have been devoted to improving the generation quality of single-concept personalization, whereas the multi-concept strategy that presents multiple concepts in a single image is even more challenging. While numerous existing multi-concept schemes focus on identity preservation of multiple concepts [22, 29], more recent approaches [13, 30] point to the importance of conceptual scalability. Nevertheless, these methods [2, 24, 39] frequently encounter challenges related to identity degradation in the absence of centralized fusion tuning, thereby hindering the plug-and-play realization. Moreover, the entangled representation incorporated in the above methods prevents users from local fine-grained control while maintaining consistency over the generated images. Notably, text-to-image diffusion models integrate text tokens through cross-attention, and several studies [15, 26] reveal the semantic-disentangled property of this mechanism, as illustrated in Figure 2. In search of more robust representations, some personalization works [28, 48] also explore the capacity of disentangled representation learning within a single diffusion model. Unfortunately, they overlook the inherent properties of robust text semantic and spatial alignment in diffusion models, consequently compromising scalability.

Given the scalability and fine-grained control of personalized concepts, we pose a question: *is it possible to derive content by compositing disentangled semantics across multiple diffusion models centrally?* To answer this question, we propose **D^{2}Plan** that learns a semantic-**d**isentangled diffusion space to **d**erive compositional content for controllable spatial-temporal **p**erson**aliza**tio**n**. Specifically, we propose Instance-Level Spatial Personalization (ISP) with the Disentangled Low-Rank Adaptation (DisLoRA) that encapsulates the individual concept, compositing semantics centrally across multiple diffusion models in the disentangled space for plug-and-play scalability. Following the gradient-guided decomposition of global and sub-prompts, ISP also facilitates local fine-grained semantic control of the generated image by adjusting the sub-prompt, while preserving consistency with the unchanged global prompt. Based on ISP, we further introduce Event-Level Temporal Personalization (ETP) to interpolate the keyframes generated by the ISP according to the time-varying prompt template decomposed by LLM, thereby enabling video personalization of specific instances. In addition, our proposed framework is compatible with ControlNet [47], which supports more conditions (*e.g.*, skeleton and depth) for improving the controllability of spatial-temporal personalization.

To summarize, our main contributions are as follows:

- We propose Instance-Level Spatial Personalization (ISP) integrated with the Disentangled Low-Rank Adaptation (DisLoRA), achieving plug-and-play scalability of multi-concept personalization in the disentangled space.
- We design a cooperative mechanism between global and sub-prompts to composite the fine-grained semantics while maintaining consistency, which is derived from the gradient-guided decomposition of prompts.
- We extend to present Event-Level Temporal Personalization (ETP) for video personalization, accomplished through the interpolating keyframes generated by the ISP based on time-varying templates decomposed by LLM.

2. Related Works

2.1. Diffusion-based Personalized Generation

Given several images of the target concept, personalization for diffusion models is typically achieved by concept-specific inversion in the text embedding space [11, 13, 29, 40, 48] or fine-tuning a pre-trained diffusion model [6, 13, 22, 28, 30, 35]. Building on earlier efforts [11, 35, 40] that focus on a single specific concept, multi-concept approaches [13, 22, 28, 30, 48] strive to incorporate various new concepts within an image, accompanied by a textual description that covers multiple concepts. In addition to the image personalization research, recent studies [5, 6] begin to generate temporally consistent customized videos that faithfully preserve the visual features of the multiple given concepts. In this study, we are dedicated to acquiring a disentangled semantic space to guarantee scalability and fine-grained control in spatial-temporal personalization, which is distinct from the previous works (as outlined in Table 1).

CVPR #9666

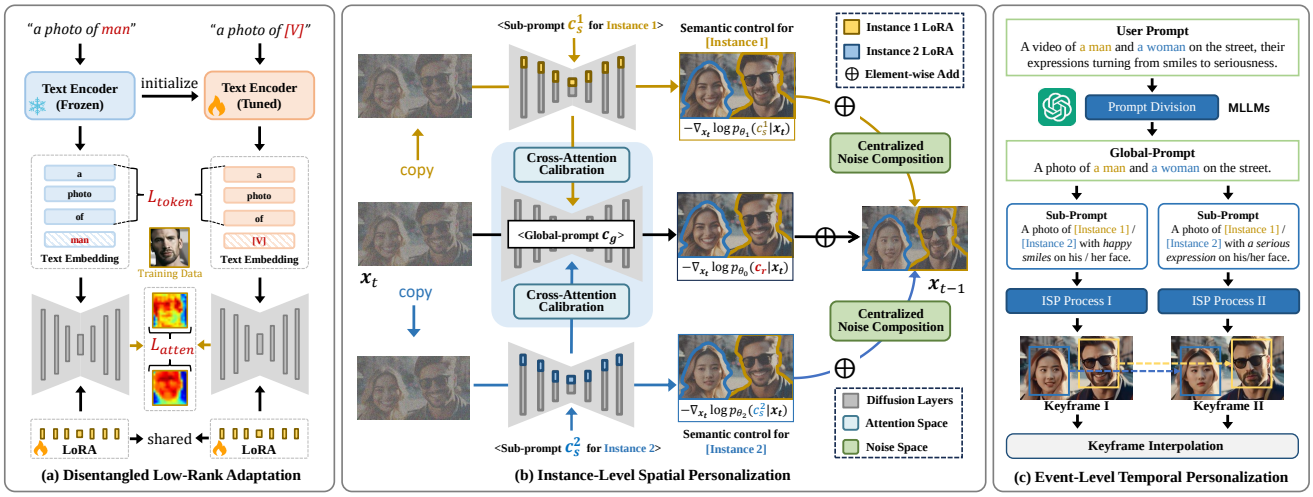


Figure 3. **Framework of the proposed D^{2}Plan**: (a) Disentangled Low-Rank Adaptation (DisLoRA) encapsulates each concept individually, facilitating the plug-and-play scalability of concepts; (b) Instance-Level Spatial Personalization (ISP) composites the concepts centrally with a cooperative mechanism between global and sub-prompts, which allows local fine-grained control under consistency; and (c) Event-Level Temporal Personalization (ETP) applies time-varying prompt templates divided by LLM for video keyframe interpolation.

2.2. Disentangled Space of Diffusion Models

Unlike the generative models of Variational AutoEncoders (VAE) [17, 44], the latent variables of diffusion models (DMs) lack explicit semantic meaning [32]. Therefore, recent works engage in the exploration of semantic representation spaces and the identification of disentanglement factors. Specifically, a series of techniques [32, 41, 45, 46] learn the semantic-disentangled space of images in an unsupervised way by introducing an extra learnable encoder, while treating DM as a conditional decoder to reconstruct images based on the disentangled semantics. A study more related to our research is MoA [28], which explores an attention space with disentangled subject-context representation for customized generation. However, MoA diminishes the multi-concept scalability due to replacing the original self-attention mechanism in the diffusion model. Besides the disentanglement representation learning for DMs, several studies [15, 26] reveal the inherent property of the disentangled semantic space within the cross-attention maps of a single text-to-image model, indicating the DMs are capable of incorporating the disentangled words into the generation process. Motivated by this discovery, we perform a centralized composition of personalized semantics encapsulated independently by the proposed DisLoRA across the multiple diffusion models in the disentangled space.

3. Preliminary

Diffusion Models (DMs) [19, 38] are generative models that learn to gradually denoise from a Gaussian distribution to a specific data distribution. Generally, given a sample x_0, the forward process gradually corrupts it into an approximate standard Gaussian distribution x_T by a predefined T-step scheduler. In the reverse process, the denoiser ϵ_θ (*e.g.*, a U-Net) is trained to reduce noise, aiming to restore the data distribution gradually. To improve the computational efficiency, Latent Diffusion Models (LDMs) perform the diffusion process in the latent space constructed by the Variational AutoEncoder (VAE). In the text-to-image setting, the generated data is expected to conform to textual cues c. Then, the training objective can be simplified as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(x_t, c, t)\|_2^2], \quad (1)$$

where ϵ denotes the noise sampled from the Gaussian distribution. During the inference, the DMs start with the initial noise $x_T \sim \mathcal{N}(0,1)$ and generate results following the conditional gradient guidance $\sigma_t \nabla_{x_t} \log p_\theta(c|x_t)$ [19], where σ_t is the hyperparameter predefined by the type of sampler. In Classifier Guidance (CG) [10], $p_\theta(c|x_t)$ is a classifier, while Classifier-Free Guidance (CFG) [18] implicitly treat it as $\nabla_{x_t} \log p_\theta(c|x_t) = -\frac{1}{\sigma_t}[\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t)]$. Then the sampling process of CFG in noise space is presented as:

$$\hat{\epsilon}_\theta(x_t, c) = (1 - w) \cdot \epsilon_\theta(x_t) + w \cdot \epsilon_\theta(x_t, c), \quad (2)$$

where w is the guidance weight for conditional and unconditional proportions, balancing fidelity and diversity [10]. Note that we omit the step t in denoiser ϵ_θ for simplicity.

4. Methodology

We present our proposed spatial-temporal personalization framework known as D^{2}Plan in Figure 3, which learns a semantic-disentangled space and composites scalable concepts. The D^{2}Plan comprises the following components: (a) DisLoRA encapsulates each personalized concept independently within the semantic-disentangled space of diffu-

CVPR #9666

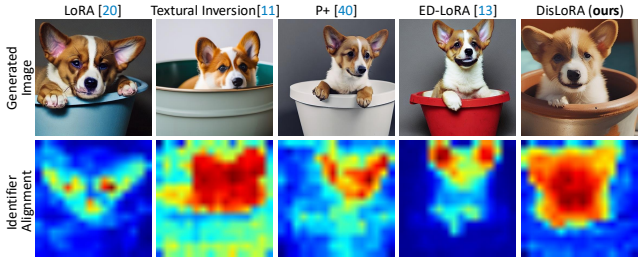


Figure 4. **Comparison with Single-Concept Methods**. Our DisLoRA shows superior disentangled alignment in attention space, enabling the subsequent composition of plug-and-play concepts.

sion models (Section 4.1); (b) Instance-Level Spatial Personalization (ISP) achieves fine-grained composition under consistency with no-degraded multi-concept scalability, derived from cooperation mechanism between global and sub-prompts (Section 4.2); and (c) Event-Level Temporal Personalization (ETP) enables video personalization by interpolating keyframes prompted by the time-varying ISP template organized by LLM (Section 4.3).

4.1. Disentangled Low-Rank Adaptation

Despite the inherent properties of robust text semantic and spatial alignment in diffusion models, the disentanglement of identifiers is weakened in single-concept personalization, as illustrated in Figure 4. Therefore, the existing single-concept methods are inadequate for plug-and-play multi-concept scalability. Motivated by previous works [15, 26], we propose DisLoRA presented in Figure 3 (a) to disentangle the new concepts. Specifically, we initially add the constraint for vanilla LoRA [20] at the token level:

$$\mathcal{L}_{token} = \mathbb{E}_{c_1,c_2}[\sum_{i=1}^{n-1} \|E_t(c_1)_i - E_f(c_2)_i\|_2^2], \quad (3)$$

where c_1 and c_2 are prompts with class token and identifier, respectively, and n is the number of tokens. E_t and E_f denote tuned and frozen text encoders, respectively. Here we assume that the identifier is at the end of the text prompt. $\mathcal{L}_{token}$ promotes the optimization process to pay more attention to personalized concepts, thus effectively avoiding the concept conflicts in plug-and-play composition. Subsequently, we impose the constraint at the attention level:

$$\mathcal{L}_{atten} = \mathbb{E}_{A_1,A_2}[1 - \frac{A_1(n) \cap A_2(n)}{A_1(n) \cup A_2(n)}], \quad (4)$$

where A_1 and A_2 are cross-attention maps of class token and identifier, respectively. $\mathcal{L}_{atten}$ encourages alignment between identifier and class token in attention space, achieving through the Intersection over Union (IoU) loss. Then the total loss is formed by L_{LDM}, L_{token}, and L_{atten}. The disentangled enhancement of DisLoRA encapsulates each concept individually, facilitating the plug-and-play composition of scalable concepts. We then describe the scalable multi-concept composition in Section 4.2.

4.2. Instance-Level Spatial Personalization

Based on DisLoRA which encapsulates each personalized concept in the disentangled space, we propose Instance-Level Spatial Personalization (ISP) for scalable multi-concept composition, as presented in Figure 3 (b). In contrast to previous multi-concept methods [13, 22], ISP could efficiently fuse non-degraded concepts without any parameter tuning. The ISP model θ aggregates identity semantics from multiple sub-diffusion models θ_i, $i \in \{1, \ldots, m\}$ equipped with DisLoRA into a global diffusion model θ_0, thus facilitating plug-and-play scalability. Within the centralized structure, we design global- and sub-prompts for the global- and sub-models, which control the overall layout and local semantics, thus maintaining consistent content across a series of generated images. Subsequently, we present a semantic cooperation mechanism that relies on the gradient-guided decomposition of the global and sub-prompts, which consists of two components: Cross-Attention Calibration and Centralized Noise Composition.

Cooperation of Global and Sub-Prompts. ISP utilizes a collaborative strategy with a global prompt c_g and m sub-prompts c_s^i, $i \in \{1, \ldots, m\}$. The former prompt c_g determines the global spatial layout, while the latter sub-prompt set c_s controls the local fine-grained personalized semantics. As exemplified in Figure 3 (b), the global prompt "*a photo of actor and actress*" provides the instance categories in the scene, and the sub-prompt "*[V] actor is smiling*" for the DisLoRA-tuned sub-model describes the personalized instance detailedly with the unique identifier. This cooperation is achieved by modifying the gradient-guided condition ($-\nabla_{x_t} \log p_\theta(c|x_t) \rightarrow -\nabla_{x_t} \log p_\theta(c_g, c_s|x_t)$) since the disentangled properties of discrete words in text-to-image diffusion models [46], which can be further expressed as:

$$\begin{aligned} &\nabla_{x_t} \log p_\theta(c_g, c_s|x_t) \\ &= \underbrace{\nabla_{x_t} \log p_\theta(c_g|c_s, x_t)}_{(a)\ global} + \underbrace{\nabla_{x_t} \log p_\theta(c_s|x_t)}_{(b)\ sub}, \end{aligned} \quad (5)$$

where (a) means the guidance of original global gradient condition c_g influenced by personalized condition c_s, and (b) denotes the guidance given the sub-prompts c_s. Then $\nabla_{x_t} \log p_\theta(c_g|c_s, x_t)$ and $\nabla_{x_t} \log p_\theta(c_s|x_t)$ drive the implementation of our centralized fusion mechanism.

Cross-Attention Calibration. As mentioned above, global condition c_g is influenced by fine-grained personalized condition c_s, thus we could introduce a new condition c_r to represent this effect, as depicted in a red letter in Figure 3 (b). Therefore, we re-formula the global gradient guidance term $\nabla_{x_t} \log p_\theta(c_g|c_s, x_t)$ in Equation (5) as:

$$\nabla_{x_t} \log p_\theta(c_g|c_s, x_t) \equiv \nabla_{x_t} \log p_\theta(c_r|x_t). \quad (6)$$

The interaction (*i.e.*, the effect of c_r) of global and sub-prompts is expected to control local fine-grained semantics.

To this end, we instantiate such effect of c_r with cross-attention map manipulation in the semantic-disentangled space [15, 26, 46]. We propose a Cross-Attention Calibration (CAC) strategy to embody the impact of each sub-prompt. In detail, our proposed CAC integrates the specific semantics derived from individual sub-prompts c_s^i across diverse models into the global prompt c_g as follows:

$$\mathrm{A}_r(k) = \begin{cases} \mathrm{A}_s(j), & \text{if } c_g(k) \text{ is class token.} \\ \mathrm{A}_g(k), & \text{otherwise} \end{cases} \tag{7}$$

where A_g and A_s represent the cross-attention map of global and sub-prompt, respectively. A_r denotes manipulated cross-attention map, k denotes the token index of the global prompt, and j represents the index of the unique identifier in c_s^i that corresponds to the class token.

Centralized Noise Composition. Recent studies [12, 25] reveal that latent noise significantly affects scene layout and enables concept composition at the individual instance level. Thus, we propose the Centralized Noise Composition (CNC) to rearrange the noise portion of each instance appropriately. Recall that we apply the DisLoRA for multiple models to encapsulate diverse personalized concepts independently. CNC uniformly schedules noise across models to generate consistent results, which instantiates $\nabla_{x_t} \log p_\theta(c_g, c_s|x_t)$ in the noise space. Specifically, c_s contains m sub-prompts that are assumed to be independent of each other. Then the gradient flow influenced by sub-prompts can be expressed as:

$$\nabla_{x_t} \log p_\theta(c_s|x_t) = \sum\nolimits_{i=1}^{m} \nabla_{x_t} \log p_{\theta_i}(c_s^i|x_t), \tag{8}$$

therefore, we regard the effect of Equation (5) as the noise blending process of newly influenced condition c_r and sub-prompts c_s. To activate the specific noise for personalized instances and maintain scene consistency, we apply a mask set $M_i,\ i \in \{0, \ldots, m\}$ for the CNC sampling process:

$$\begin{aligned} \hat{\epsilon}_\theta(x_t, c) = (1-w)\cdot\epsilon_{\theta_0}(x_t) + w\cdot M_0\cdot\epsilon_{\theta_0}(x_t, c_r) \\ + w\cdot\sum\nolimits_{i=1}^{m} M_i\cdot\epsilon_{\theta_i}(x_t, c_s^i), \end{aligned} \tag{9}$$

where $M_i,\ i \in \{1, \ldots, m\}$ is the instance masks that are automatically obtained from cross-attention map according to specific token [26, 27] and $M_0 = \mathbf{1} - \sum_{i=1}^{m} M_i$. This process essentially involves rearranging term $\epsilon_\theta(x_t, c)$ in Equation (2). Practically, we can also generate the image with the global prompt and use Detectron2 [43] and EfficientViT-SAM [49] to expedite the generation of instance masks. Then x_{t-1} is predicted from x_t using $\hat{\epsilon}_\theta$ and is copied $m+1$ times to input into global and sub-models in the next step.

4.3. Event-Level Temporal Personalization

ISP achieves plug-and-play scalability for multi-concepts personalization, which controls local semantics while ensuring consistency across a series of generated images. Consequently, we propose Event-Level Temporal Personalization (ETP) to implement video personalization based on keyframe interpolation, as shown in Figure 3 (c). Unlike previous keyframe-based text-to-video generation [8, 50], we apply the LLM to execute temporal planning at the event level and achieve video personalization of specific instances. Specifically, given a text prompt c describing activities involving multiple instances, we utilize LLM to decompose it into global prompt c_g and time-varying sub-prompts$\{c_s^1, c_s^L\}$ that match ISP for keyframe interpolation:

$$F_1, F_2, \cdots, F_L = \texttt{ETP}(\texttt{ISP}(c_g, c_s^1), \texttt{ISP}(c_g, c_s^L)), \tag{10}$$

where F denotes the personalized frames and L is the sequence length. Here we integrate the off-the-shelf interpolation method (*i.e.*, SEINE [8]) into $\texttt{ETP}(\cdot)$ for video personalization. We utilize some in-context examples with LLM to reason out the strategy for dividing sub-prompts of the event. Please refer to **Supplementary Material** for the detailed template of our instructions. The global prompt and sub-prompts obtained by the above reasoning facilitate the generation of temporally consistent keyframes.

5. Experiments

5.1. Implementation Details and Evaluation Setup

Implementation Details. We implement our D^{2}Plan with the pre-trained Stable-Diffusion-XL[1]. DisLoRA integrates the LoRA layer [20] with a rank of 128 into the linear layer of all attention modules. We train the LoRA parameters using the AdamW optimizer [37] with a constant learning rate of 3×10^{-4} for U-Net and 3×10^{-6} for text encoder. All experiments are conducted on a single NVIDIA A100 GPU with 80GB of memory. During inference, we use the 50-step EulerDiscreteScheduler [21] with a CFG [18] scale of 7.5. For the coherence and stability of generation, we execute ISP from $t \leq 35$ to avoid instance conflicts by the prior guidance of the pre-trained model. The weights of L_{token} and L_{atten} are set to 1×10^{-3} and 1×10^{-2}, respectively.

Dataset. Following previous works [5, 6], we collect a dataset that contains 12 concepts to evaluate our proposed D^{2}Plan. This dataset comprises different categories of concepts (including characters and objects), each containing 16 unique images of the target concept in different contexts.

Evaluation Metrics. Following prior works [13, 40], we evaluate our method using *Image Alignment* in the CLIP image feature space [11], which measures the similarity between generated and reference images. Also, we apply *Text Alignment* to measure the text-image similarity (*i.e.*, CLIP score [16]) between generated images and given prompts. Additionally, we adopt ArcFace score [9] for *Identity Alignment*, which illustrates the identity-preserving capabilities

[1] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

Method	Text Alignment ↑			Image Alignment ↑			Identity Alignment ↑			Consistency ↑		
	Single	Multi	Δ	Single	Multi	Δ	Single	Multi	Δ	Single	Multi	Δ
DreamBooth [35]	0.695	0.688	−0.007	0.770	0.681	−0.089	0.631	0.640	+0.009	0.792	0.624	−0.168
Textual Inversion [11]	0.683	0.685	+0.002	0.767	0.696	−0.077	0.538	0.538	+0.000	0.653	0.577	−0.076
P+ [40]	0.710	0.705	−0.005	0.754	0.748	−0.006	0.569	0.564	−0.005	0.686	0.583	−00103
Custom Diffusion [22]	0.672	0.720	+0.048	0.750	0.722	−0.048	0.605	0.550	−0.055	0.709	0.591	−0.118
Mix-of-show [13]	0.700	0.685	−0.015	0.763	0.748	−0.015	0.606	0.608	+0.002	0.757	0.655	−0.102
ISP+LoRA (**ours**)	0.712	0.729	+0.017	0.765	0.758	−0.007	0.678	0.660	−0.018	0.836	0.813	−0.023
ISP+DisLoRA (**ours**)	**0.730**	**0.737**	+0.007	**0.773**	**0.765**	−0.008	**0.705**	**0.692**	+0.013	**0.837**	**0.826**	−0.011

Table 2. **Quantitative Results for Spatial Personalization**. We report the results of different methods with single- and multi-concept, and calculate the metric difference Δ to represent the performance gap between the single- and multi-concept. The best results are **bolded**.

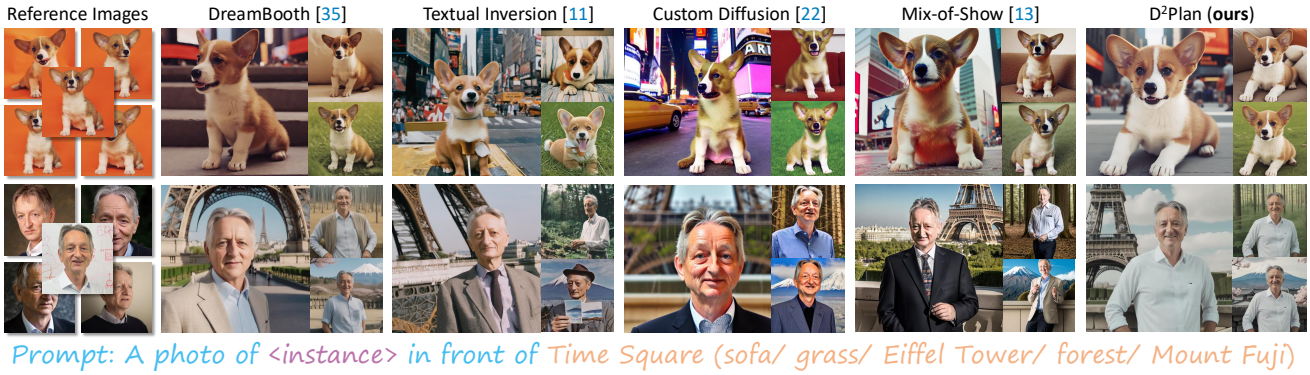


Figure 5. **Qualitative Comparisons of *Single-Concept Spatial* Personalization**. In object and character personalization, our D^{2}Plan control scene while maintaining identity consistency, which outperforms other compared methods in both controllability and quality.

Method	Identity Alignment↑			Consistency ↑		
	Single	Multi	Δ	Single	Multi	Δ
DreamBooth [35]	0.822	0.725	−0.097	0.510	0.528	+0.018
Textual Inversion [11]	0.749	0.658	−0.091	0.485	0.480	−0.015
P+ [40]	0.785	0.604	−0.181	0.504	0.520	+0.016
Custom Diffusion [22]	0.708	0.725	+0.017	0.488	0.476	−0.012
Mix-of-show [13]	0.793	0.685	−0.108	0.558	0.634	+0.076
D^{2}Plan (**ours**)	**0.832**	**0.818**	−0.014	**0.900**	**0.932**	+0.032

Table 3. **User Study for Spatial Personalization**. Participants evaluate generation quality from aspects of single-/multi-identity preservation and local fine-grained control under consistency.

Method	DINO ↑	CLIP-T ↑	T-Cons ↑	Human ↑
DB [35]+AD [14]	0.280	0.221	0.938	0.611
CD [22]+AD [14]	0.289	0.236	0.911	0.630
VideoDreamer [5]	0.362	0.225	0.948	0.658
DisenStudio [6]	0.391	0.247	0.963	0.663
D^{2}Plan (**ours**)	**0.663**	**0.559**	**0.979**	**0.812**

Table 4. **Quantitative Results for Temporal Personalization.** We compare our D^{2}Plan with the image personalization methods (*i.e.*, DreamBooth (DB) [35] and Custom Diffusion (CD) [22]) by integrating AnimateDiff (AD) [14]. Human denotes normalized ratings by participants in the user study, including visual confusion, attribute binding, subject missing, and action binding.

by detecting the target identity in generated images. We further calculate the T-Cons score [42] to quantify the *Consistency* of a series of locally controlled images with the same global prompt. Besides, we apply the DINO score [6, 35] to measure temporal identity preservation, calculating the DINO feature cosine similarity between the generated and given image. In addition to the above objective metrics, we hire 32 participants on Amazon Mechanical Turk [3] to conduct *User Study*. With these metrics, we compare the D^{2}Plan with recent state-of-the-art personalization methods for image (*i.e.*, DreamBooth [35], Textual Inversion [11], P+ [40], Custom Diffusion [32], and Mix-of-Show [13]) and video (*i.e.*, VideoDreamer [5] and DisenStudio [6]).

5.2. Quantitative Comparisons

Spatial Personalization. In Table 2, we present the comparison results for single- and multi-concept spatial personalization. Our D^{2}Plan outperforms other methods in generic evaluated metrics, revealing its superior performance. In the realm of multi-concept personalization, our D^{2}Plan achieves an Image Alignment score of 0.765 and an Identity Alignment score of 0.692, showcasing an obvious improvement in identity preservation. Benefiting from the plug-and-play DisLoRA that encapsulates individual concepts in the disentangled space, our D^{2}Plan also minimizes identity degradation from single to multiple concepts. Moreover, our D^{2}Plan exhibits superior fine-grained controllability, with Consistency scores of 0.837 and 0.826 for single- and multi-concept, respectively. In addition, we conduct a user study for spatial personalization, as shown in Table 3. We apply 18 prompts and generate 5 images per

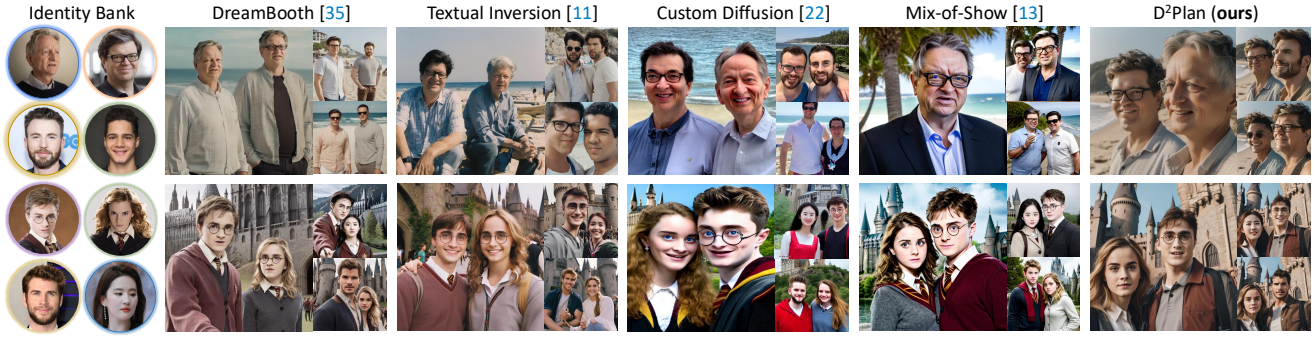


Figure 6. **Qualitative Comparisons of *Multi-Concept Spatial* Personalization**. Given multiple identities of a specific character, our D^{2}Plan controls local semantics (*i.e.*, identity, location, and appearance change following sub-prompts) of the target character while maintaining scene consistency, while other methods suffer from multi-identity preservation and conceptual confusion of the same category.

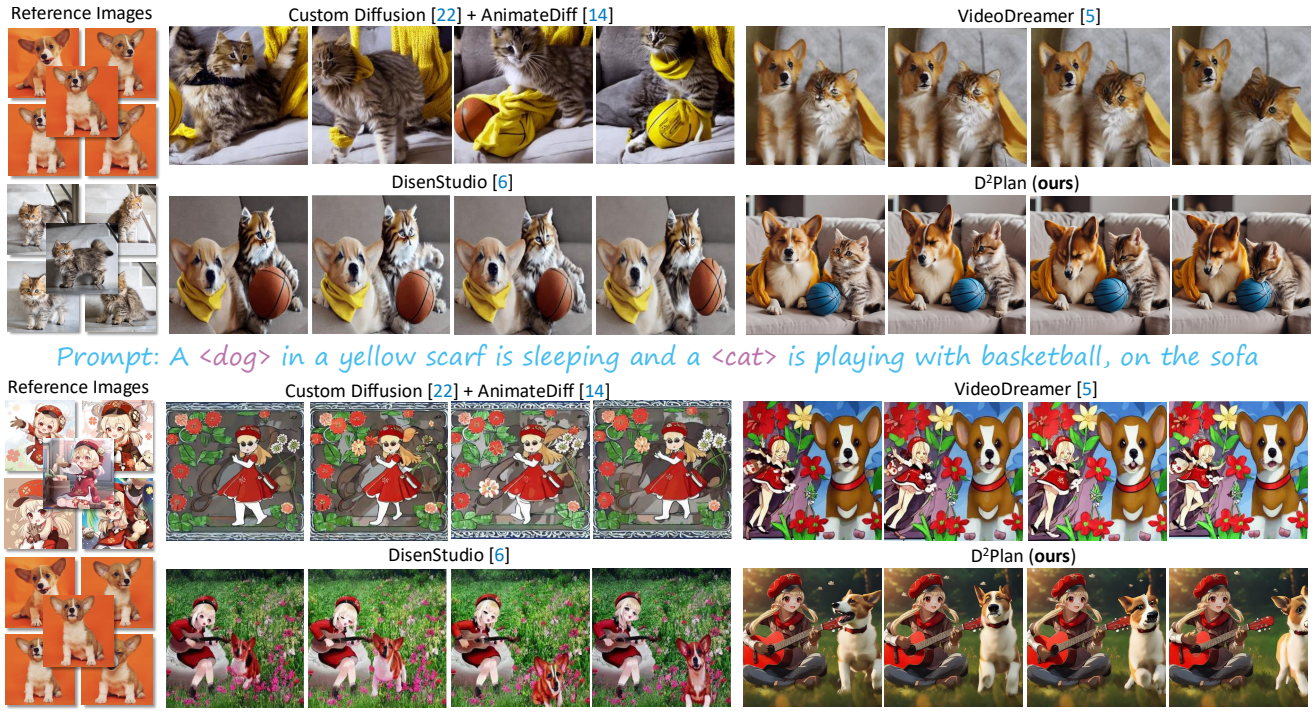


Figure 7. **Qualitative Comparisons of *Temporal* Personalization**. Benefiting from spatial consistency personalization of the ISP, our D^{2}Plan exhibits superior generation quality and temporal consistency compared to other video personalization methods. Besides, D^{2}Plan mitigates typical drawbacks in previous video methods, such as visual confusion, attribute binding, subject missing, and action binding.

prompt for each method, yielding a total of 540 images. We ask participants to score the generated images from $\{1, 2, 3, 4, 5\}$, considering aspects of identity preservation and fine-grained semantic control under consistency. Leveraging concept disentanglement and non-degraded composition with local control, our D^{2}Plan achieves the highest normalized scores for both single- and multi-concept.

Temporal Personalization. We summarize the comparison and user study for temporal personalization in Table 4. We report DINO, CLIP-T, and T-Cons scores for measuring identity preservation, text alignment, and temporal consistency, respectively. Our D^{2}Plan surpasses all compared video methods, confirming the efficacy of local controllable consistency of ISP for temporal personalization. Similar to spatial personalization, we also conduct a user study on temporal personalization, measuring visual confusion, attribute binding, subject missing, and action binding. The results show that our method is favored by most participants.

5.3. Qualitative Comparisons

Single-Concept Personalization. We showcase the qualitative comparisons of single-concept methods in Figure 5. As shown in the results, our D^{2}Plan closely adheres to the given identify and maintains excellent scene consistency for

both object and character. Taking the "*man*" in the second row as illustrations, previous methods hardly disentangle identity from scenarios (*e.g.*, "*Eiffel Tower*" and "*Mount Fuji)*"), whereas our D^{2}Plan not only preserves identity but also allows local modifications (*e.g.*, background change).

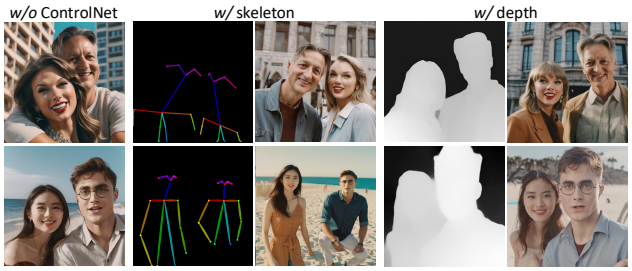
Multi-Concept Personalization. The qualitative results of multi-concept generation are shown in Figure 6. As previously mentioned, we fine-tune individual concepts with distinct DisLoRA representations and composite fine-grained semantics across multiple diffusion models in the disentangled space. Benefiting from this mechanism, we resolve ambiguity in personalized concepts with the same category. For example, compared with Mix-of-Show [13] that depends on the user-provided spatial condition to distinguish different identities, our D^{2}Plan clarifies the identities of "*[V_1] man and [V_2] man*" in the first row. In addition to the superiority of multi-identity preservation and conceptual confusion, our method effectively controls the local fine-grained semantics of the target character while ensuring scenario consistency, such as adding "*sunglasses*" in the first row and changing identity in both rows.

Temporal Personalization. In Figure 7, we provide qualitative comparisons of multi-concept video personalization. While the previous approach for multi-concept image personalization (*i.e.*, Custom Diffusion [32]) integrated with AnimateDiff [14]) struggles to preserve multiple identities and generate consistent frames, our D^{2}Plan technique addresses this challenge by interpolating consistent event-level keyframes generated by ISP. This demonstrates the effectiveness of our spatial consistency generation for temporal personalization. Compared to specialized video personalization methods [5, 6], our D^{2}Plan significantly improves the generation quality and alleviates visual confusion. For example, we disentangle the in-domain visual confusion of DisenStudio (*i.e.*, "*dog*" and "*cat*" in the first example) and cross-domain visual confusion of VideoDreamer (*i.e.*, "*girl*" and "*dog*" in the second example). Furthermore, by regulating the user-provided prompt into the time-varying sub-prompts that control the local semantics of each instance, we prevent attribute and action binding, exemplified by the "*yellow scarf*" in the first example and "*dog is running*" in the second example.

5.4. Ablation study

Ablation for ISP. In Figure 8 (a), we gradually investigate the contribution of each module in ISP to controllability, which is exemplified by scene variation of a single concept for clarity. We observe that the CAC and CNC collaborate to boost both identify preservation and consistency maintenance. Taking the samples in the first row converting the "*girl*" from the "*Eiffel Tower*" to the "*Hogwarts Castle*" as illustrations, we find that turning off CAC leads to identity degradation, while scene consistency is perfectly protected.

(a) Ablation for Instance-Level Spatial Personalization (ISP).



(b) Ablation for additional spatial conditions.

Figure 8. **Ablation Studies of D^{2}Plan**. The above result encompasses: (a) Effect of Cross-Attention Calibration (CAC), Centralized Noise Composition (CNC), and number of instances N; and (b) Effect of additional spatial conditions, *i.e.*, skeleton and depth.

The overall layout changes significantly when CNC is discarded, indicating that CNC dominates consistency maintenance. In addition, we present the results with more instances in the last column, which illustrates the effectiveness of the ISP in personalizing diverse concepts.

Ablation for additional conditions. We also assess practicality by adding extra spatial conditions (*i.e.*, skeleton and depth). Figure 8 (b) presents the results of our D^{2}Plan integrated with ControlNet [47]. It is worth noting that additional spatial conditions are optional for the multi-concept generation of our D^{2}Plan, but are necessary for previous methods such as Mix-of-Show [13], otherwise instances of the same category may be confused (as shown in Figure 6).

6. Conclusion

In this work, we leverage the semantic-disentangled diffusion space to composite fine-grained semantics, aiming to generate controllable personalized content at the spatial and temporal levels. Concretely, we first propose Instance-Level Spatial Personalization with the Disentangled Low-Rank Adaptation for plug-and-play multi-concept personalization, which achieves local fine-grained control by the decomposed gradient guidance of global and sub-prompts. Then, we introduce Event-level Temporal Personalization for multi-instance temporal synthesis by interpolating ISP-generated keyframes, where an LLM is utilized to divide temporal-varying ISP prompts. Extensive experiments demonstrate that our framework outperforms recent state-of-the-art methods qualitatively and quantitatively.

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, page 8, 2023. 1

[2] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19650–19661, 2023. 2

[3] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, pages 3–5, 2011. 6

[4] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 1

[5] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*, 2023. 2, 5, 6, 8

[6] Hong Chen, Xin Wang, Yipeng Zhang, Yuwei Zhou, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. *arXiv preprint arXiv:2405.12796*, 2024. 2, 5, 6, 8

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 1

[8] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 5

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 5

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8780–8794, 2021. 1, 3

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2, 5, 6

[12] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22930–22941, 2023. 5

[13] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 4, 5, 6, 8

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 6, 8

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 3, 4, 5

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5

[17] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 3

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 5

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 1, 3

[20] 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. *arXiv preprint arXiv:2106.09685*, 2021. 4, 5

[21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 26565–26577, 2022. 5

[22] 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, 4, 6

[23] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 1

[24] Haonan Lin. Dreamsalon: A staged diffusion framework for preserving identity-context in editable face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8589–8598, 2024. 2

[25] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composi-

tion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2294–2305, 2023. 5

[26] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. *arXiv preprint arXiv:2401.11739*, 2024. 2, 3, 4, 5

[27] Trong-Tung Nguyen, Duc-Anh Nguyen, Anh Tran, and Cuong Pham. Flexedit: Flexible and controllable diffusion-based object-centric image editing. *arXiv preprint arXiv:2403.18605*, 2024. 5

[28] Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, Kfir Aberman, et al. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. *arXiv preprint arXiv:2404.11565*, 2024. 2, 3

[29] Lianyu Pang, Jian Yin, Haoran Xie, Qiping Wang, Qing Li, and Xudong Mao. Cross initialization for face personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8393–8403, 2024. 2

[30] 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, 2024. 2

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[32] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10619–10629, 2022. 3, 6, 8

[33] 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. 1

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2

[35] 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. 1, 2, 6

[36] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[37] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4596–4604, 2018. 5

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[39] Maitreya Suin, Nithin Gopalakrishnan Nair, Chun Pong Lau, Vishal M Patel, and Rama Chellappa. Diffuse and restore: A region-adaptive diffusion model for identity-preserving blind face restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6343–6352, 2024. 2

[40] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2, 5, 6

[41] Yunnan Wang, Ziqiang Li, Zequn Zhang, Wenyao Zhang, Baao Xie, Xihui Liu, Wenjun Zeng, and Xin Jin. Scene graph disentanglement and composition for generalizable complex image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[42] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7623–7633, 2023. 6

[43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[44] Baao Xie, Qiuyu Chen, Yunnan Wang, Zequn Zhang, Xin Jin, and Wenjun Zeng. Graph-based unsupervised disentangled representation learning via multimodal large language models. *Advances in Neural Information Processing Systems (NeurIPS)9*, 2024. 3

[45] Tao Yang, Yuwang Wang, Yan Lv, and Nanning Zheng. Disdiff: Unsupervised disentanglement of diffusion probabilistic models. *arXiv preprint arXiv:2301.13721*, 2023. 3

[46] Tao Yang, Cuiling Lan, Yan Lu, et al. Diffusion model with cross attention as an inductive bias for disentanglement. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3, 4, 5

[47] 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 (ICCV)*, pages 3836–3847, 2023. 2, 8

[48] Yanbing Zhang, Mengping Yang, Qin Zhou, and Zhe Wang. Attention calibration for disentangled text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4764–4774, 2024. 2

[49] Zhuoyang Zhang, Han Cai, and Song Han. Efficientvit-sam: Accelerated segment anything model without performance loss. *arXiv preprint arXiv:2402.05008*, 2024. 5

[50] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 5