UniversalBooth: Model-Agnostic Personalized Text-to-Image Generation

Songhua Liu[1,2], Ruonan Yu[1], and Xinchao Wang[1*]
[1]National University of Singapore [2]School of Artificial Intelligence, Shanghai Jiao Tong University
{songhua.liu,ruonan}@u.nus.edu, xinchao@nus.edu.sg

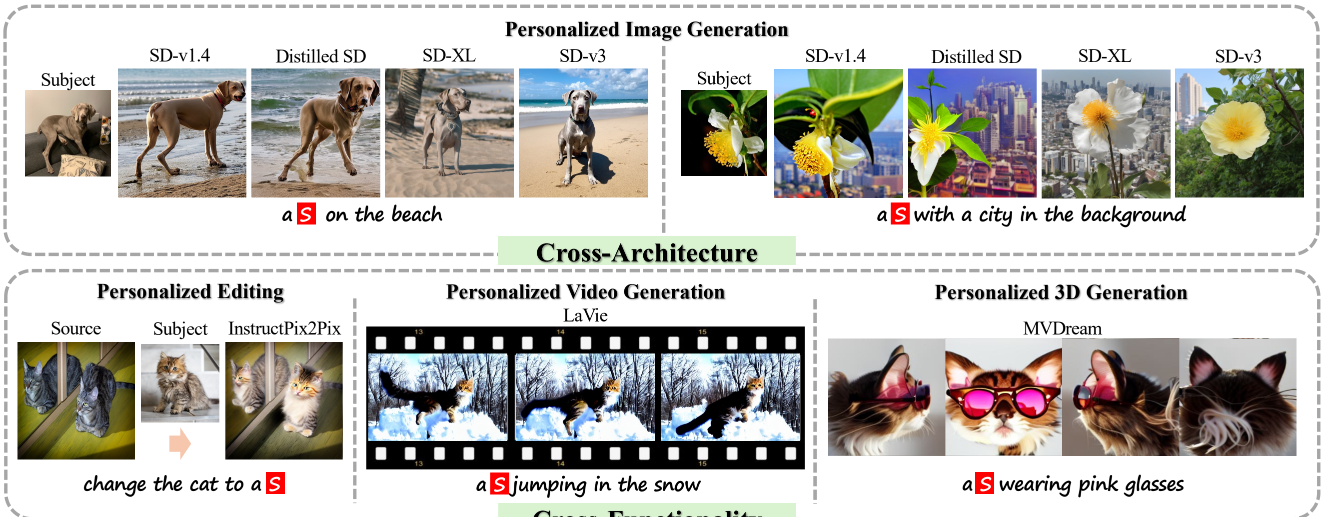

Figure 1. We propose a model-agnostic personalized text-to-image generation method termed UniversalBooth. Once trained, it can be applied to various diffusion models with different architectures and functionalities without any additional training. "S" is a virtual token referring to the input subject. The seen model during training is SD-v1.4, while the unseen models are Distilled SD [21], SD-XL [32], SD-v3-Medium [13], InstructPix2Pix [5], LaVie [46], and MVDream [40].

Abstract

Given a source image, personalized text-to-image generation produces images preserving the identity and appearance while following the text prompts. Existing methods heavily rely on test-time optimization to achieve this customization. Although some recent works are dedicated to zero-shot personalization, they still require re-training when applied to different text-to-image diffusion models. In this paper, we instead propose a model-agnostic personalized method termed UniversalBooth. *At the heart of our approach lies a novel cross-attention mechanism, where different blocks in the same diffusion scale share common square transformation matrices of key and value. In this way, the image encoder is decoupled from the diffusion architecture while maintaining its effectiveness. Moreover, the cross-attention performs hierarchically: the holistic attention first captures the global semantics of user inputs for textual combination with editing prompts, and the fine-grained attention divides the holistic attention scores for various local patches to enhance appearance consistency. To improve the performance when deployed on unseen diffusion models, we further devise an optimal transport prior to the model and encourage the attention scores allocated by cross-attention to fulfill the optimal transport constraint. Experiments demonstrate that our personalized generation model can be generalized to unseen text-to-image diffusion models with a wide spectrum of architectures and functionalities without any additional optimization, while other methods cannot. Meanwhile, it achieves comparable zero-shot personalization performance on seen architectures with existing works.*

1. Introduction

Although text-to-image diffusion models have advanced significantly in recent years due to their generative capabilities [11, 31, 36], they fall short in personalized image generation, also known as subject-driven image generation, where generated images adhere to text prompts while preserving specific identities and appearances from user-

*Corresponding Author.

provided images. Given the broad applicability in real-world scenarios, this customization has garnered attention from both academia and industry.

Personalized generation typically involves establishing correspondences between source images and textual space, allowing diffusion models to reconstruct images and generate variations based on text prompts. Early methods rely on test-time optimization [14] or fine-tuning the diffusion model [22, 37], which are computationally expensive and impractical for end users. Recent efforts aim to reduce this burden through offline learning, such as training a visual encoder for feed-forward textual correspondences [19, 24, 39, 44, 47, 52]. These methods achieve impressive zero-shot personalized text-to-image generation: once trained, the visual encoder can capture input concepts from subject images in real time.

However, flexibility remains an issue when applying these methods to various text-to-image models with unseen structures, which is a practical problem since real-world models can be updated or replaced frequently. Re-training the encoder for each model can take several days on multiple GPUs [44, 47], making it highly cumbersome if not impossible at all. Moreover, for some distilled models like LCM [28], it is even infeasible to conduct vanilla training directly as it requires a specific distillation objective concerning its teacher.

Focusing on this drawback, we aim at a model-agnostic approach termed *UniversalBooth* in this paper and expect a trained visual encoder to be generalized to other text-to-image backbones seamlessly. To this end, we first delve into the design intricacies of existing personalized generation methods without test-time optimization and reveal that the significant impediment to such generalization lies in the strong coupling between the visual encoder and the cross-attention layers of the diffusion UNet, which are key modules for the diffusion model to interact with input conditions. Previous works largely ignore the variability in backbone text-to-image models. Consequently, their visual encoders are susceptible to overfitting the specific diffusion model seen in training.

In this paper, we tackle this challenge by introducing a novel cross-attention mechanism that decouples the visual subject encoder from the diffusion architecture. Unlike conventional approaches, our method utilizes a shared group of square mapping matrices for both key and value components across different blocks within the same diffusion scale. This design promotes flexibility across a range of diffusion models characterized by varying numbers of channels or blocks, all while ensuring consistent effectiveness in generating high-quality images.

Moreover, we devise a hierarchical cross-attention strategy. Specifically, the visual encoder first capsules global semantic features of user inputs into a virtual word, which is convenient for textual combination with editing prompts. To enhance appearance consistency, the cross-attention scores for this word are further divided by fine-grained attention considering local patches. Overall, this hierarchical approach ensures that the generated images not only align with textual prompts but also maintain coherence and fidelity to the visual subjects.

Notably, the fine-grained cross-attention in our approach is backed up with an optimal transport prior, which is achieved by regulating attention scores allocated by the cross-attention mechanism to fulfill optimal transport constraints. In this way, even when tested on unfamiliar novel diffusion architectures, our customization model can be guided by this prior knowledge acquired in training, which further bolsters cross-model generalizability.

We conduct extensive experiments to showcase the effectiveness and versatility of our approach. As shown in Fig. 1, results indicate that once trained, the proposed UniversalBooth can be generalized to unseen text-to-image diffusion models with a wide spectrum of architectures and functionalities without any additional training effort, while other methods cannot. Meanwhile, it yields on-par zero-shot personalized generation performance with existing works on seen diffusion backbones. Our contributions can be summarized as follows:

- We investigate a novel problem, namely model-agnostic personalized text-to-image generation. To the best of our knowledge, this is the first work dedicated to the cross-model generalization issue in this field.
- We tailor a novel cross-attention mechanism to address the problem. Specifically, it adopts shared key and value mappings among various blocks within the same scale, works in an innovative hierarchical manner, and is injected with optimal transport prior.
- Experiments suggest that UniversalBooth achieves superior cross-model personalized generation results to unseen diffusion models and comparable performance in the vanilla test setting on seen models.

2. Related Works

Personalized text-to-image generation refers to producing images according to the text prompts while preserving the identity and appearance of users' image inputs. One promising solution for this application is to learn the word embedding [14] or fine-tune the diffusion model [22, 37] specifically for one subject in an iterative optimization fashion, which exhibits limited flexibility. Recent studies have been dedicated to addressing this limitation and focused on an any-subject-one-model paradigm. The basic idea is to replace the optimization process with a single feed-forward propagation: to learn a neural network and map the input images to the conditional space of the diffusion model [9, 16, 18, 24, 26, 29, 41, 43, 44, 47, 50, 52] in a one-

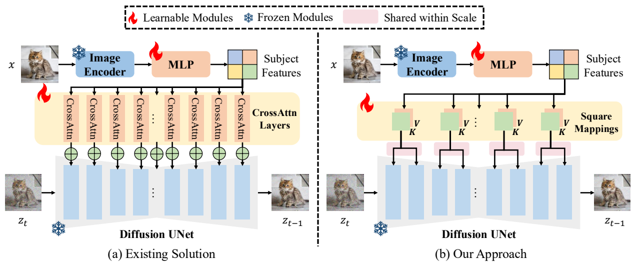

Figure 2. (a) Overview of zero-shot personalized text-to-image generation solutions (text branch omitted). (b) Our approach learns square and shared key-value mappings in cross-attention layers, enabling flexible cross-diffusion-model generalization during inference without extra training.

stop manner. We provide a systematic summary of existing works in the appendix.

These methods typically focus on *the subject-wise flexibility* and have achieved remarkable performance simultaneously. Nevertheless, when the base diffusion model changes—a common and practical scenario given the rapid proliferation and advancement of large text-to-image diffusion models [3, 4, 10, 12, 13, 20, 21, 28–30, 32, 34, 34, 36, 38, 42, 44, 48]—they typically require adapting the subject mapper to the new architecture through an optimization process as well [35]. We thus offer a different perspective regarding *the model-wise flexibility* and introduce UniversalBooth in this paper, the first method specifically designed for the cross-model generalization problem in personalized generation. Unlike existing techniques, UniversalBooth enables zero-shot customized generation on unseen diffusion models without requiring any additional training.

3. Methodology

3.1. Preliminary

Different from early approaches [14, 22, 37] that obtain textual correspondences of subject images through test-time optimization, recent test-time fine-tuning-free personalized image generation methods learn a neural mapping from the visual space to the textual space so that the textual representations can be generated in a single forward propagation.

We illustrate the overall pipeline of some popular designs like ELITE [47] and IP-Adapter [50] in Fig. 2(a). Typically, given a subject image x, these methods first adopt the CLIP image encoder [33], denoted as $\phi(\cdot)$, for feature extraction. As CLIP has been trained on abundant text-image pairs to align corresponding features, it may serve as a valuable resource for learning text-aware vision representations. Subsequently, a learnable MLP denoted as M is trained to convert CLIP vision features into virtual words in the textual embedding space. Additional cross-attention is incorporated into the cross-attention layers of the pre-trained diffusion UNet to enhance its adaptability to

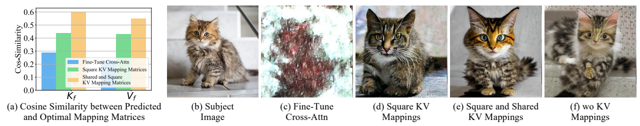

Figure 3. Preliminary results of cross-model generalization by different operations on cross-attention. Images in (c), (d), (e), and (f) are generated with the prompt `A photo of a S`.

conditions injected by subject images, which is crucial to the performance given the significant role of cross-attention (CA) in subject-driven text-to-image generation, as highlighted in [1, 15, 22]. The CA results for the subject branch are added to the original results:

$$\text{Out} \leftarrow \text{Out} + \lambda \text{CA}(Q, (M \circ \phi(x))\hat{W}_k, (M \circ \phi(x))\hat{W}_v), \tag{1}$$

where Q is the query matrix mapped from features in the diffusion backbone, $\hat{W}_k$ and $\hat{W}_v$ are learnable key and value mappings respectively, and λ is a hyperparameter controlling the strength of subject injection.

We assume that Stable Diffusion v1.5 [36] is adopted here. It first learns an auto-encoder $(\mathcal{E}(\cdot), \mathcal{D}(\cdot))$, where the encoder $\mathcal{E}(\cdot)$ maps an image x to a lower dimensional latent space: $z \leftarrow \mathcal{E}(x)$, and the decoder $\mathcal{D}(\cdot)$ learns to decode z back to the image space $\hat{x} \leftarrow \mathcal{D}(z)$ such that $\hat{x}$ is close to the original x. Denoising is conducted in the latent space by a UNet $\epsilon_\theta(\cdot)$ for noise prediction. With a pre-trained and frozen auto-encoder, text encoder, and UNet, the MLP and additional cross-attention layers are optimized using the vanilla noise prediction loss $\mathcal{L}_{simple}$ [17, 31, 36]:

$$\mathcal{L}_{simple} = \mathbb{E}_{z,y,\epsilon,t}[\|\epsilon - \epsilon_\theta(z_t, t, \tau(y), M \circ \phi(x))\|_2^2], \tag{2}$$

where $\tau(\cdot)$ represents the text encoder, y denotes the text input, t is the denoising step, and z_t is the latent codes at step t. Typically, the text y is drawn from some templates, such as `A photo of S`, where `S` can be instantiated as virtual words [47] or the corresponding class name [50]. In the inference time, `S` can be flexibly composed with natural languages to achieve customization.

3.2. Shared and Square KV Mapping Matrices

This paper aims at a plug-and-play customization model that enables users to utilize a diverse range of diffusion models with varying structures during testing. Achieving this goal necessitates ensuring that the customization encoder remains independent of the architecture of the diffusion UNet. However, recall the pipeline of existing techniques shown in Fig. 2(a), and we find that the encoder and the UNet are coupled in the cross-attention layers. When using a different architecture, the number of cross-attention layers and their dimensions do not necessarily match the seen model. Consequently, the trained customized model cannot be loaded into a novel diffusion model. To address this issue, instead of fine-tuning key and value mappings for cross-attention, we propose to learn additional square transformation matrices T_k and T_v for features after the image MLP, such that their shapes are only relevant to the image feature dimension, without any dependence on diffusion backbones. The key and value mappings used in the added cross-attention layers are obtained via linear transformation: $\hat{W}_k \leftarrow T_k W_k$ and $\hat{W}_v \leftarrow T_v W_v$.

We find that such a technique kills two birds with one stone: it not only resolves the issue of variability in the number of channels across various models, but also reduces the upper bound of generalized error even if their channel dimensions are consistent. Please refer to the theoretical analysis in the appendix, which indicates that our approach is less sensitive to the discrepancy of feature spaces between seen and unseen models. In Fig. 3(a), we validate this effect by comparing the cosine similarity between the estimated $\hat{W}_k$ and $\hat{W}_v$ in unseen architectures and their optimal counterparts that have been trained on these architectures and serve as oracles. As shown in Figs. 3(c) and (d), this design is essential for the model to produce meaningful results.

At this point, the only unresolved issue is the variability in the number of cross-attention layers. To tackle the problem, it is crucial to discern between the invariant and variant factors across different target diffusion models. By ensuring that our method does not rely on variant factors, we can develop a solution that remains robust and adaptable across various diffusion models. In this paper, we capitalize on the multi-scale functionality inherent in diffusion UNets and introduce an innovative solution whereby the cross-attention layers within different blocks but the same resolution scale share a common set of T_k and T_v matrices. Formally, the cross-attention results for the j-th block of the i-th scale can be written as:

$$\mathrm{CrossAttn}(Q^{i,j}, (M \circ \phi(x)) T_k^i W_k^{i,j}, (M \circ \phi(x)) T_v^i W_v^{i,j}), \tag{3}$$

where ^{i,j} specifies the indices of scale and block. The overall design is illustrated in Fig. 2(b). As shown in Figs. 3(d) and (e), this strategy further enhances the cross-model generalizability.

One might wonder whether it is feasible to retain T_k and T_v as identity matrices and solely focus on learning the MLP part, yielding the simplest design. However, comparing Figs. 3(e) and (f), the method results in inferior identity preservation, underscoring the significance of adapting the textual condition space to the subject condition space. Please refer to the appendix for more explorations.

3.3. Hierarchical Cross-Attention

To address the trade-off between appearance preservation and text prompt adherence, we devise a novel hierarchical cross-attention mechanism in this paper, where the holistic attention first captures the global semantics of subject images, and then the attention scores to the global semantics

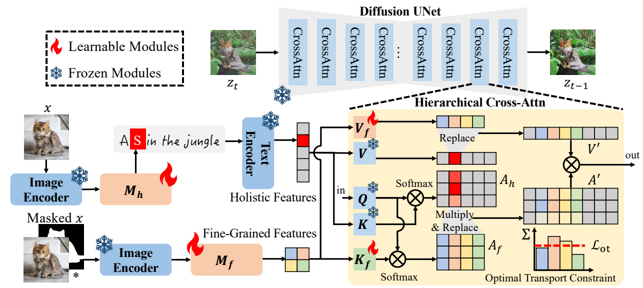

Figure 4. The proposed personalized text-to-image generation method is driven by holistic and fine-grained mappings. We devise a hierarchical cross-attention mechanism for the interaction between the two branches, which yields satisfactory text adherence and appearance preservation concurrently. An optimal transport constraint is applied here as prior knowledge guiding cross-model generalization.

are divided according to the fine-grained attention. Similar to ELITE [47], we subsequently trained two mapping networks M_h and M_f to extract holistic and fine-grained features in the textual space for holistic and fine-grained cross-attention, respectively.

Holistic Mapping: The designs of the holistic mapping network M_h mainly follow the previous work BLIP-Diffusion [24], which extracts n virtual words in the token embedding space of the CLIP text encoder via a Q-Former [25], taking as input features from a frozen pre-trained image encoder. Denoting these virtual words as `S`, the Q-Former is optimized to minimize a loss function $\mathcal{L}_h$ consisting of $\mathcal{L}_{simple}$ in Eq. 2 based on the text prompts like `A photo of S` and an L1 regularization term for `S`. Since the primary goal of holistic mapping is to extract virtual words that are compatible with real prompts, we omit T_k and T_v in Eq. 3, and instead directly use the native key-value parameters to process the textual conditions.

Fine-Grained Mapping: The designs of the fine-grained mapping network M_f mainly follow the previous work ELITE [47], where features of n layers in the CLIP image encoder are separately mapped to n virtual words in the token embedding space of the CLIP text encoder by n learnable sub-mappers with two `Linear-LayerNorm-LeakyReLU` blocks. Following ELITE [47], we also apply subject masks here to filter out irrelevant backgrounds.

Different from previous techniques that require adjusting the hyperparameter λ in Eq. 1 to balance the holistic and fine-grained attention, we propose a hierarchical approach. Denoting the query, key, and value after the projection of W_q, W_k, and W_v in a holistic cross-attention layer as $Q \in \mathbb{R}^{s \times c}$ and $K_h, V_h \in \mathbb{R}^{l_h \times c}$, respectively, where s is the number of query tokens, *i.e.*, the spatial dimensionality of current latent codes, l_h is the number of textual tokens, and c is the feature dimensionality of this layer, and assuming that the primary word is located at

Method	Unseen Architecture												Seen Architecture		
	LCM			Base Diffusion			Small Diffusion			Tiny Diffusion					
	C-T	C-I	D-I	C-T	C-I	D-I	C-T	C-I	D-I	C-T	C-I	D-I	C-T	C-I	D-I
BLIP-Diffusion [24]	.282	.589	.201	.275	.680	.201	.266	.652	.360	.281	.565	.181	.300	.771	.583
IP-Adapter [50]	.291	.679	.449	.281	.596	.210	.269	.538	.133	.268	.536	.142	.295	**.796**	.629
ELITE [47]	.288	.665	.441	.219	.535	.050	.229	.559	.093	.226	.562	.121	.255	.762	.652
Ours	**.307**	**.694**	**.532**	**.303**	**.710**	**.500**	**.302**	**.687**	**.482**	**.300**	**.675**	**.479**	**.302**	**.772**	**.667**

Table 1. Quantitative comparisons with state-of-the-art zero-shot text-to-image personalization methods and ablation studies. Best performance is marked in **bold**.

the p-th token, we first compute the holistic attention map A_h with $A_h \leftarrow \text{Softmax}(\frac{QK_h^\top}{\sqrt{c}})$, and extract the attention map corresponding to the primary word, *i.e.*, the p-th column of A_h, denoted as $A_h^{:,p} \in \mathbb{R}^{s\times 1}$. Then, denoting the post-projection key and value in the corresponding fine-grained cross-attention layer as $K_f, V_f \in \mathbb{R}^{l_f\times c}$, respectively, where l_f is the number of fine-grained tokens, we further divide the column $A_h^{:,p}$ into l_f columns weighted by the fine-grained attention map $A_f \leftarrow \text{Softmax}(\frac{QK_f^\top}{\sqrt{c}})$, *i.e.*, replace the column $A_h^{:,p}$ by l_f columns $A_h^{:,p} * A_f$, with $*$ representing element-wise multiplication allowing broadcast. Accordingly, the p-th row in V_h is substituted by the l_f rows in V_f. Denoting the updated attention map and value as A' and V', respectively, the output of such hierarchical cross-attention is given by $\text{Out} \leftarrow A'V'$, which is adopted to replace the original computational rule in Eq. 1 for all cross-attention layers. We offer an illustrative presentation of the workflow in Fig. 4.

3.4. Optimal Transport Prior

Without knowledge of unseen models, subject-driven models face challenges in achieving cross-model generalization. To mitigate this issue, we seek to imbue fine-grained attention with generic knowledge, guiding the cross-model customization process with priors to enhance performance. In this paper, we explore an optimal transport prior that encourages an even migration of visual patterns from subject images to customized results and penalizes one-to-many mappings [27]. Assume that $Q \in \mathbb{R}^{s\times c}$ and $K_f \in \mathbb{R}^{l_f\times c}$ are two discrete distributions. The total mass in K is defined as the total attention score to the primary word in the holistic attention, *i.e.*, $\sum A_h^{:,p}$. Since we expect the mass in K to be evenly transported into Q, the mass of each point in K should be $\frac{\sum A_h^{:,p}}{l_f}$. We add the regularization of balanced total attention scores in each point of K to Eq. 2. The loss function for the fine-grained mapping can be written as:

$$\mathcal{L}_f = \mathcal{L}_{simple} + \frac{\alpha}{N}\sum_{j=1}^{N} \| \sum A_f^{:,j} - \frac{\sum A_h^{:,p}}{l_f} \|_2^2, \quad (4)$$

where N denotes the total number of columns in all the fine-grained attention maps, and α is a hyperparameter controlling the strength of this regularization.

4. Experiments

4.1. Implementation Details

We build UniversalBooth on the open-source implementation of ELITE [47]. The architecture of the diffusion model in training is StableDiffusion v1.4 [36]. The test set of the OpenImages dataset [23], containing 120K images, is adopted to train the holistic mapping, while 47K of them with annotations of object masks are used to train the fine-grained mapping. The hyper-parameter α in Eq. 4 is set as 0.01 empirically. We train the holistic and fine-grained mappings on 4 RTX 6000 Ada GPUs for $40,000$ and $80,000$ iterations, respectively. Other setups, including the diffusion sampler and the scale of classifier-free guidance, follow the default configurations if not mentioned specifically.

The evaluation dataset is also consistent with ELITE [47] that adopts $2,500$ test cases formed by a pairwise combination of 20 subject images, 25 text templates, and 5 random seeds. Following the convention of personalized text-to-image generation [22, 37, 47], we evaluate our method on 3 metrics, including `CLIP-I` (C-I) and `DINO-I` (D-I) for image consistency and `CLIP-T` (C-T) for text consistency. `CLIP-I` and `DINO-I` measure the cosine similarity between features of generated and source subject images in the CLIP image encoder [33] and the ViTS/16 DINO [6]. `CLIP-T` measures the cosine similarity between features of generated images in the CLIP image encoder and text prompts in the CLIP text encoder, where object class names of the subject images are used in the text templates.

4.2. Cross-Architecture Generalization

As our target in this paper is a model-agnostic personalized text-to-image generation method, we mainly evaluated the proposed UniversalBooth on diffusion models with unseen architectures in training. Specifically, we consider both cases of large-to-small and small-to-large cross-model generalization. For small architectures, we adopt two diffusion models distilled from Stable Diffusion in [21] for experiments, denoted as *Small Diffusion* and *Tiny Diffusion*. For large architectures, we consider two popular structures, *i.e.*, *StableDiffusion-XL* [32] (SD-XL) and *StableDiffusion-3-Medium [13]* (SD-3). Although both seen and unseen involve CLIP for textual embedding, for SD-XL and SD-

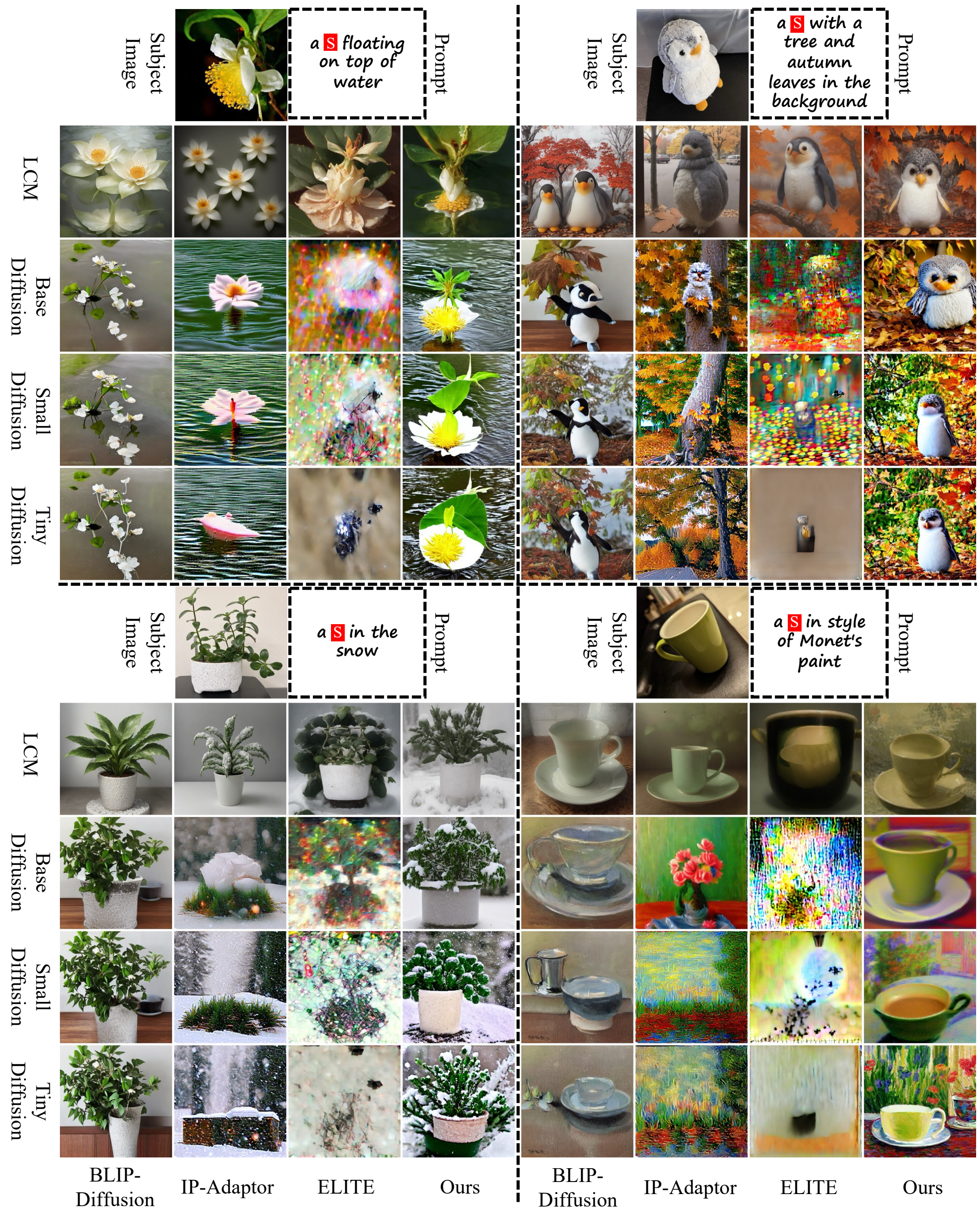

Figure 5. Results of large-to-small cross-architecture generalization comparing with state-of-the-art methods.

3, since the conditional spaces are constructed using multiple text encoders, they do not strictly satisfy the consistent condition space assumption required by the fine-grained encoder. Therefore, we use only the holistic encoder for them.

To demonstrate the unique superiority of our method in cross-architecture generalization, we compare Universal-Booth with three open-source and state-of-the-art personalized generation methods, including BLIP-Diffusion [24], IP-Adapter [50], and ELITE [47]. These methods overlook the variability of diffusion models in the inference time and couple the subject encoder module with the diffusion UNet. For instance, BLIP-Diffusion requires fine-tuning both the whole diffusion UNet and the image-to-BLIP mapping module. IP-Adapter and ELITE fine-tune the cross-attention layers in the original diffusion backbones, which makes the subject encoders overfit to the seen diffusion ar-

Figure 6. Results of small-to-large cross-architecture generalization. The seen architecture is StableDiffusion v1.4, while the unseen architectures are *StableDiffusion-XL* and *StableDiffusion 3 Medium*, respectively.

Architecture	Method	CLIP-T	CLIP-I	DINO-I
SD-XL	ELITE	**.317**	.650	.336
	Ours	.294	**.766**	**.580**
SD-3-Medium	ELITE	.316	.691	.446
	Ours	**.319**	**.750**	**.581**

Table 2. Quantitative comparisons with the baseline method ELITE [47] on small-to-large cross-architecture generalization. Best performance is marked in **bold**.

chitecture in training. To adapt them for cross-architecture generation, we only load the matched cross-attention layers in inference time and drop the extra layers.

As a result, as shown in Fig. 5, without specific consideration of cross-model generalization, existing methods fail to produce plausible personalized text-to-image generation results. For BLIP-Diffusion, since it utilizes BLIP [25] as a pre-trained vision-language prior, the produced results can often convey aligned semantics. However, the colors and textures cannot match those in the original subject images. Even worse, when the architectural gap between the unseen and seen diffusion models is large, *e.g.*, Tiny Diffusion, the content layouts tend to be out of control. For IP-Adapter and ELITE, suffering from misalignments of feature spaces in this setting, they are prone to messy and meaningless textures. Compared with them, our method successfully addresses these and exhibits better robustness to the architectural variations. Quantitatively, as reported in Tab. 1(left), our method outperforms existing ones in cross-architecture generalization by a large margin.

For small-to-large generalization with SD-XL and SD-3, the qualitative and quantitative comparisons against the ELITE baseline [47] are shown in Fig. 6 and Tab. 2, respectively. Although ELITE can capture the overall semantics of subject images and textual prompts, the results largely overlook the detailed appearances. In contrast, our method exhibits superior cross-model generalization.

4.3. Comparisons on Seen Architectures

We also compare our method with existing ones on the seen architectures. As shown in Tab. 1(right), UniversalBooth

Figure 7. Results of personalized generation on seen architectures.

overall achieves comparable quantitative metrics with state-of-the-art methods. Specifically, it yields higher `CLIP-T` and `DINO-I` but slightly lower `CLIP-I`. We speculate that it is because the hierarchical cross-attention mechanism in this paper improves the trade-off between text alignment and the preservation of detailed local patterns, which may favor low-level metrics like `DINO-I` based on features of self-supervised learning compared with the high-level metric `CLIP-T`. Also, according to the ablation studies, the sharing of key and value mappings in cross-attention layers of the same scale inevitably sacrifices performance on seen architectures to some extent. In addition, compared with works like BLIP-Diffusion and IP-Adapter, the consumption of computational resources, including data, GPU cards, and training time, is significantly lower for our method, as demonstrated in the appendix. These are factors that UniversalBooth has not achieved significantly superior performance to the state-of-the-art methods on seen architectures.

Nevertheless, as shown in Fig. 7, our method indeed has

Figure 8. Existing methods like ELITE [47] require adjusting the hyperparameter λ to balance the text adherence and appearance preservation, whose optimal choices, marked by the stars, are case by case, while our method does not.

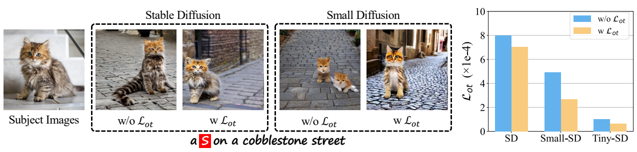

Figure 9. The optimal transport prior $\mathcal{L}_{ot}$ helps cross-model generalization by regulating the layout of the generated subject.

significant advantages in many cases. Specifically, methods like BLIP-Diffusion and IP-Adapter tend to preserve the major semantics while ignoring the local patterns, which leads to inferior appearance preservation, *e.g.*, the cat, the bag, and the cat statue in the 1st, 4th, and 5th columns, respectively. ELITE struggles to deal with colors in text prompts by confusing colors of subjects and backgrounds, *e.g.*, the dog in the 3rd column, and the bag in the 4th column. In comparison, our method handles these cases better.

4.4. Ablation Studies

In this part, we demonstrate the effectiveness of the proposed three components, including shared and square key and value mapping matrices, hierarchical cross-attention, and optimal transport prior, through ablation studies. Quantitatively, the impact of each component on the performance is shown in Tab. 3.

Shared and Square Key and Value Mapping Matrices: As demonstrated in Fig. 3, by decoupling the subject encoder from the number of channels and blocks in diffusion backbones, shared and square key and value mappings help align the subject feature space and the condition space shared by different diffusion models and play crucial roles in achieving cross-architecture generalization. We provide more supportive examples in the appendix.

Hierarchical Cross-Attention: Previous methods like IP-Adapter and ELITE achieve the trade-off between text prompt adherence and appearance preservation by adjusting the weight of the subject condition, which is not robust in practice. As shown in Fig. 8, in different cases, the optimal choice of this hyperparameter can also be different. Compared with these methods, the hierarchical cross-attention proposed in this paper can balance the two worlds better without deliberate adjustment.

Setting	Unseen Architecture						Seen Architecture		
	Small Diffusion			Tiny Diffusion					
	C-T	C-I	D-I	C-T	C-I	D-I	C-T	C-I	D-I
w/o KV	**.302**	.664	.424	**.300**	.653	.422	.301	.756	.650
w/o Shared KV	.283	.650	.365	.232	.556	.163	.300	**.785**	**.683**
w/o HieraAttn	.289	.614	.291	.288	.613	.305	.297	.768	.629
w/o $\mathcal{L}_{ot}$	.292	.644	.377	.284	.613	.291	**.302**	.760	.584
Ours	**.302**	**.687**	**.482**	**.300**	**.675**	**.479**	**.302**	.772	.667

Table 3. Ablation studies for various technical designs introduced in this paper. `HieraAttn` denotes the hierarchical attention. Best performance is marked in **bold**.

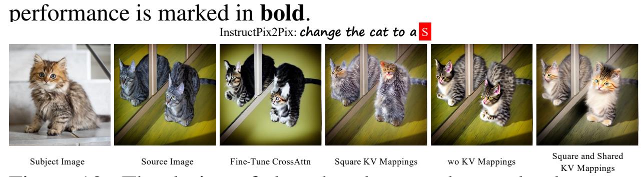

Figure 10. The design of shared and square key and value mappings enables UniversalBooth to be generalized to models with various functionalities like text-driven editing.

Optimal Transport Prior: The optimal transport prior is useful to regulate the semantic layout of the generated results by penalizing one-to-many mappings. We illustrate this effect in Fig. 9. Quantitative measurements on the test cases also validate that adding the regularization $\mathcal{L}_{ot}$ in training would lead to lower $\mathcal{L}_{ot}$ in cross-model inference.

4.5. Further Extension

Interestingly, we find that the proposed UniversalBooth is also compatible with other text-to-image models with different functionalities, like InstructPix2Pix [5] for text-driven image editing, thanks to the universal subject space resulting from the shared and square key and value mapping matrices. As shown in Fig. 10, the model would migrate incorrect or insufficient subject patterns if square mappings or shared mappings are removed.

5. Conclusions

In this paper, we present UniversalBooth, a model-agnostic framework for personalized text-to-image generation. Dedicated to the problem of cross-diffusion generalization, we mainly introduce three novel designs: (1) cross-attention with shared and square key and value mappings, which achieves cross-architecture zero-shot inference in functionality, (2) hierarchical cross-attention, which alleviates the trade-off between text adherence and appearance preservation, and (3) an optimal transport prior injected into the fine-grained attention, which guides the behavior of unseen models with generic knowledge and further improves the cross-model generalization performance. Experiments demonstrate that UniversalBooth is the first versatile model for personalized text-to-image generation that can be generalized to unseen diffusion backbones seamlessly without any additional training effort. It also achieves superior zero-shot personalization performance on seen architectures.

Acknowledgments

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).

References

[1] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 3

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 13

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3

[4] James Betker, Gabriel Goh, Li Jing, TimBrooks, Jianfeng Wang, Linjie Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Aditya Ramesh. Improving image generation with better captions. 2023. 3

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 8

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[7] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023. 13

[8] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 13

[9] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12501–12511, 2025. 2, 13

[10] Siying Cui, Jia Guo, Xiang An, Jiankang Deng, Yongle Zhao, Xinyu Wei, and Ziyong Feng. Idadapter: Learning mixed features for tuning-free personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 950–959, 2024. 3

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[12] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 3

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 3, 5

[14] 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. 2, 3, 13

[15] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 3, 13

[16] Jing He, Haodong Li, Yongzhe Hu, Guibao Shen, Yingjie Cai, Weichao Qiu, and Ying-Cong Chen. Disenvisioner: Disentangled and enriched visual prompt for customized image generation. *arXiv preprint arXiv:2410.02067*, 2, 2024. 2, 13

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[18] Linyan Huang, Haonan Lin, Yanning Zhou, and Kaiwen Xiao. Flexip: Dynamic control of preservation and personality for customized image generation. *arXiv preprint arXiv:2504.07405*, 2025. 2, 13

[19] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 2

[20] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: Architecturally compressed stable diffusion for efficient text-to-image generation. *ICML Workshop on Efficient Systems for Foundation Models (ES-FoMo)*, 2023. 3

[21] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. On architectural compression of text-to-image diffusion models. *arXiv preprint arXiv:2305.15798*, 2023. 1, 3, 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*, pages 1931–1941, 2023. 2, 3, 5, 13

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *Interna-*

tional Journal of Computer Vision, 128(7):1956–1981, 2020. 5

[24] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4, 5, 6, 13, 14

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 4, 7

[26] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023. 2, 13

[27] Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang. Dynast: Dynamic sparse transformer for exemplar-guided image generation. In *European Conference on Computer Vision*, pages 72–90. Springer, 2022. 5

[28] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2, 3

[29] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 2, 13

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1, 3

[32] 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, 3, 5

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5, 12

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[35] Lingmin Ran, Xiaodong Cun, Jia-Wei Liu, Rui Zhao, Song Zijie, Xintao Wang, Jussi Keppo, and Mike Zheng Shou. X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8775–8784, 2024. 3

[36] 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*, pages 10684–10695, 2022. 1, 3, 5

[37] 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*, pages 22500–22510, 2023. 2, 3, 5, 13

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[39] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2, 13

[40] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1

[41] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation. *arXiv preprint arXiv:2404.05674*, 2024. 2, 13

[42] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. 3

[43] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. 2025. 2, 13

[44] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 3, 13

[45] Qinghe Wang, Xu Jia, Xiaomin Li, Taiqing Li, Liqian Ma, Yunzhi Zhuge, and Huchuan Lu. Stableidentity: Inserting anybody into anywhere at first sight. *arXiv preprint arXiv:2401.15975*, 2024. 13

[46] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1

[47] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2, 3, 4, 5, 6, 7, 8, 13, 14

[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*, pages 1–20, 2024. 3

[49] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio:

Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023. 13

[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. 2023. 2, 3, 5, 6, 13, 14

[51] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Expanded conditioning for the personalization of attribute-aware image generation. *arXiv preprint arXiv:2305.16225*, 2023. 13

[52] Yuxuan Zhang, Jiaming Liu, Yiren Song, Rui Wang, Hao Tang, Jinpeng Yu, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. *arXiv preprint arXiv:2312.16272*, 2023. 2, 13