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

Supplementary Material

In this part, we provide a taxonomic summary of related works, theoretical analysis, more experimental results, as well as limitations and future works that cannot fit into the main paper due to the page limit.

A. Taxonomic Summary of Related Works

In Tab. 4, we provide a taxonomic summary of related works on personalized text-to-image generation from three aspects, including zero-shot flexibility, domain agnosticism, and model agnosticism.

B. Theoretical Analysis

We analyze the generalized error bounds for the two strategies: fine-tuning KV mapping weights in previous methods and learning square mapping matrices in our approach, in the following theorem:

Theorem 1. *Assume that $X_t, X_i \in \mathbb{R}^c$ are c-dimensional random variables in spaces of a text encoder and an image encoder, respectively, $W^a \in \mathbb{R}^{c\times d}$ is the key (value) mapping matrix that maps X_t to the feature space of a pretrained text-to-image diffusion model a, i.e., $X_t^a = X_t W^a$, $W_{ft}^a \in \mathbb{R}^{c\times d}$ fine-tuned from W^a represents the transformation from the image space to the diffusion feature space, i.e. $X_i^a = X_i W_{ft}^a$, and $A \in \mathbb{R}^{c\times c}$ is a square matrix that transforms a given X_i to its correspondence in the text space,* i.e., *$W_{ft}^a = AW^a$. We would like to generalize the fine-tuned W_{ft}^a to an arbitrary unseen text-to-image diffusion model b with the key (value) matrix W^b.*

If we directly apply W_{ft} to compute X_i^b, i.e., $X_i^b = X_i W_{ft}^a$, the expectation of square error between X_i^b and the optimal $X_i^{b*} = X_i W_{ft}^{b*}$, where $W_{ft}^{b*} = A^* W^b$, has the following upper bound:

$$\begin{aligned}\mathbb{E}[\|X_i^b - X_i^{b*}\|_2^2] \leq \mathbb{E}[\|X_i\|_2^2](\|A\|_2^2\|W^a - W^b\|_2^2 \\ + \|W^b\|_2^2(\|A - I\|_2^2 + \|A^* - I\|_2^2)),\end{aligned} \tag{5}$$

where I denotes an identity matrix.

If we compute X_i^b via $X_i^b = X_i A W^b$, the expectation of the square error has the following upper bound:

$$\mathbb{E}[\|X_i^b - X_i^{b*}\|_2^2] \leq \mathbb{E}[\|X_i\|_2^2]\|W^b\|_2^2(\|A - I\|_2^2 + \|A^* - I\|_2^2). \tag{6}$$

Proof. For Eq. 5, if $X_i^b = X_i W_{ft}^a$, through the given conditions, we have:

$$\begin{aligned}\mathbb{E}[\|X_i^b - X_i^{b*}\|_2^2] =& \mathbb{E}[\|X_i A W^a - X_i A^* W^b\|_2^2] \\ \leq& \mathbb{E}[\|X_i\|_2^2]\|AW^a - AW^b \\ &+ AW^b - A^*W^b\|_2^2 \\ \leq& \mathbb{E}[\|X_i\|_2^2](\|AW^a - AW^b\|_2^2 \\ &+ \|AW^b - A^*W^b\|_2^2) \\ \leq& \mathbb{E}[\|X_i\|_2^2](\|A\|_2^2\|W^a - W^b\|_2^2 \\ &+ \|W^b\|_2^2\|A - I + I - A^*\|_2^2) \\ \leq& \mathbb{E}[\|X_i\|_2^2](\|A\|_2^2\|W^a - W^b\|_2^2 \\ &+ \|W^b\|_2^2(\|A - I\|_2^2 + \|A^* - I\|_2^2)),\end{aligned} \tag{7}$$

where the 1st and 3rd inequalities stem from the submultiplicative property of matrix norms, and the 2nd and 4th inequalities are established according to the triangle inequality. For Eq. 6, if $X_i^b = X_i A W^b$, the proof is the same as that for the term $\|AW^b - A^*W^b\|_2^2$ in Eq. 7:

$$\begin{aligned}\mathbb{E}[\|X_i^b - X_i^{b*}\|_2^2] =& \mathbb{E}[\|X_i A W^b - X_i A^* W^b\|_2^2] \\ \leq& \mathbb{E}[\|X_i\|_2^2]\|AW^b - A^*W^b\|_2^2 \\ \leq& \mathbb{E}[\|X_i\|_2^2]\|W^b\|_2^2(\|A - I\|_2^2 \\ &+ \|A^* - I\|_2^2).\end{aligned} \tag{8}$$

□

We can observe that the term $\|A\|_2^2\|W^a - W^b\|_2^2$ is eliminated in our method, which indicates that our approach is insensitive to the discrepancy of feature spaces across seen and unseen backbones. Given that text and image features are based on CLIP models [33], which have been trained for text-image alignment, the term $\|A^* - I\|_2^2$ can be considered small.

We also provide an illustrative visualization in Fig. 11 for the superiority of our method when generalizing to unseen models.

C. More Experimental Results

More Ablation Studies: To illustrate the effectiveness of the proposed technical methods, we provide comprehensive ablation studies in Fig. 15, including results on 4 architectures, including the seen Stable Diffusion 1.4, Base Diffusion, Small Diffusion, and Tiny Diffusion, by the full method, the method fine-tuning the key and value mappings in cross-attention, the method with only square key and value mappings without sharing the mappings within scale,

Method	Venue	Zero-Shot	Domain-Agnostic	Model-Agnostic
TextualInversion [14]	ICLR'23	✗	✓	✗
DreamBooth [37]	CVPR'23	✗	✓	✗
Customize Diffusion [22]	CVPR'23	✗	✓	✗
E4T [15]	TOG'23	✗	✓	✗
Break-A-Scene [2]	SIGGRAPH Asia'23	✗	✓	✗
ProSpect [51]	TOG'23	✗	✓	✗
DisenBooth [7]	ICLR'24	✗	✓	✗
FaceStudio [49]	Arxiv'23	✓	✗	✗
PhotoMaker [26]	CVPR'24	✓	✗	✗
InstantID [44]	ArXiv'24	✓	✗	✗
StableIdentity [45]	ArXiv'24	✓	✗	✗
ELITE [47]	ICCV'23	✓	✓	✗
BLIP-Diffusion [24]	NeurIPS'23	✓	✓	✗
SuTI [8]	NeurIPS'23	✓	✓	✗
SubjectDiffusion [29]	SIGGRAPH'24	✓	✓	✗
IP-Adapter [50]	ArXiv'23	✓	✓	✗
InstantBooth [39]	CVPR'24	✓	✓	✗
SSR-Encoder [52]	CVPR'24	✓	✓	✗
MoMA [41]	ECCV'24	✓	✓	✗
DisEnvisioner [16]	ICLR'25	✓	✓	✗
UniReal [9]	CVPR'25	✓	✓	✗
FlexIP [18]	ArXiv'25	✓	✓	✗
OminiControl [43]	ICCV'25	✓	✓	✗
Ours	ICCV'25	✓	✓	✓

Table 4. Taxonomy of personalized text-to-image generation and comparisons with previous works.

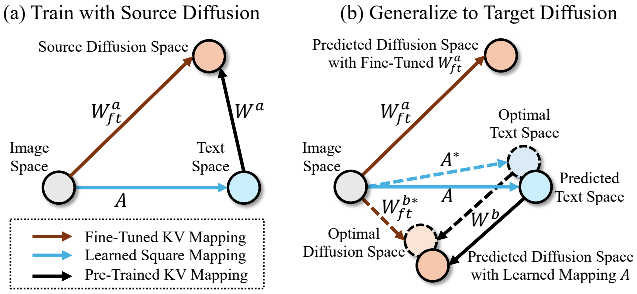

Figure 11. An illustrative visualization of previous and our method when generalizing to novel unseen diffusion models. The notions here are consistent with Theorem 1 in the main manuscript.

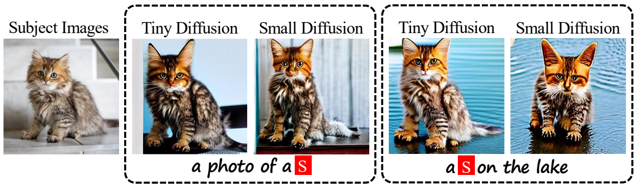

Figure 12. The shared and square key and value mappings learned by a small architecture are generalizable to larger models as well. In this example, the seen architecture is Tiny Diffusion, while the unseen model is Small Diffusion.

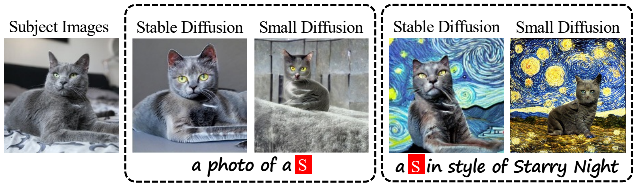

Figure 13. It is also feasible to adopt timestep-dependent subject representations to achieve similar functionalities of fine-tuning key and value mappings in existing works.

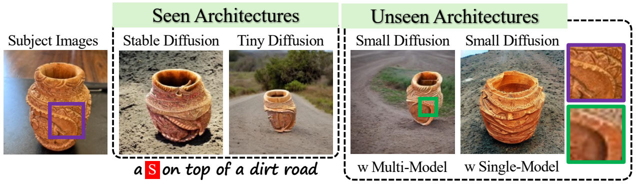

Figure 14. It would benefit cross-architecture generalization if more diffusion models are incorporated during training. In this example, we train the subject encoders on Stable Diffusion and Tiny Diffusion and test them on Small Diffusion. Compared with training only on Stable Diffusion, it performs better in the preservation of local patterns.

the method without hierarchical attention, and the method without optimal transport prior. In general, the square and shared key and value mappings are the most critical factors to achieve cross-architecture generalization. The hierarchical attention improves the preservation of local details and achieves a better trade-off between text adherence and appearance preservation. And the optimal transport prior regulates the layout of generated subjects effectively, especially on architectures with relatively large differences from the seen one, *e.g.*, Tiny Diffusion.

Generalization from Small to Large Architectures: Following existing works [24, 47, 50], we adopt Stable Diffusion as the seen architecture in training and try to generalize the subject encoder to smaller models such as Base Diffusion, Small Diffusion, and Tiny Diffusion. Here, we demonstrate that it is also feasible for small-to-large generalization. As shown in Fig. 12, we train the subject encoder and shared key and value mappings on Tiny Diffusion and

Method	GPU	Iter.	Data
BLIP-Diffusion [24]	16 A100	500K	129M
IP-Adapter [50]	8 V100	1M	10M
ELITE [47]	**4 V100**	400K	**125K**
Ours	4 RTX6000Ada	**120K**	**125K**

Table 5. Comparisons with state-of-the-art zero-shot text-to-image personalization methods on computational resources, including GPUs, the number of training iterations, and the size of training datasets. The minimum requirements are highlighted in **bold**.

test them on other architectures. The unseen models can still generate reasonable and visually appealing results.

More Designs for Better Cross-Model Generalization: As analyzed in the main paper, the key idea to achieve model-agnostic personalized text-to-image generalization is to explore shared properties across different architectures. The final choice of this paper is the shared square mappings for key and value transformation in cross-attention layers. In fact, we have also attempted other designs, like using timestep-dependent subject representations or introducing more architectures during the training time. Although the results are also encouraging, as shown in Figs. 13 and 14, we find these designs result in more complex pipelines in either inference or training time, requiring more computational resources. By contrast, the shared and square key and value mappings presented in this paper offer an overall more elegant and simple-yet-effective solution.

D. Limitations and Future Works

Future works related to the UniversalBooth proposed in this work may explore other useful common properties of various diffusion models or devise other regularizations to enhance the cross-model generalization performance. Furthermore, the generalization capability of our approach hinges on the assumption of utilizing the same text encoder, such as the CLIP text encoder. We anticipate that future research will aim to relax this constraint, thereby enabling personalized text-to-image generation to be applicable across a broader range of models with diverse textual spaces. The possible negative social impact caused by AIGC models used in this work could be potentially avoided by using specific detection methods.

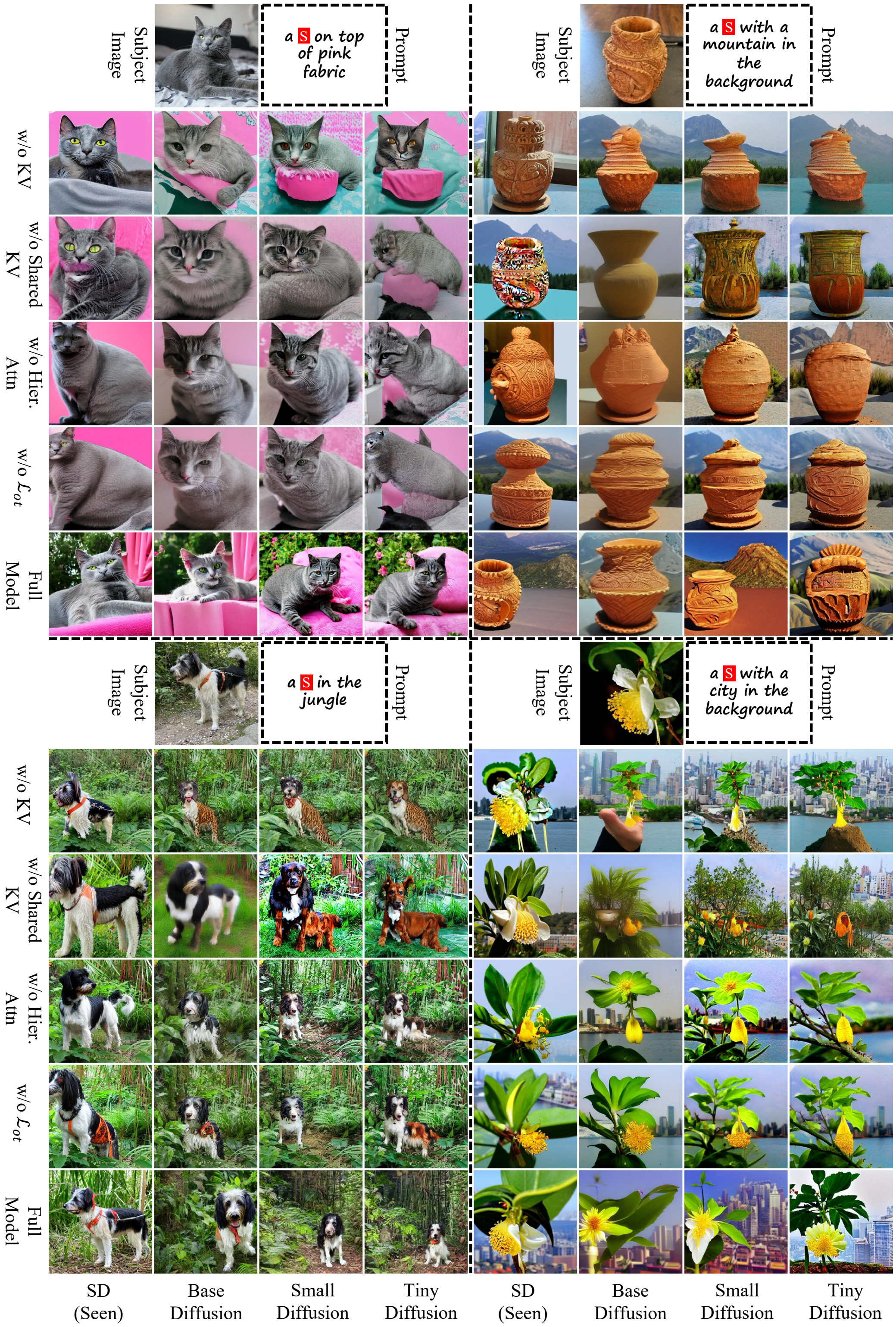

Figure 15. More ablation studies. In general, shared and square key and value mappings are the key for cross-architecture inference compared with without using additional key and value mappings or only using square mappings. Hierarchical cross-attention results in a better trade-off between text adherence and appearance preservation, and optimal transport prior helps produce more reasonable image layouts.