DCoAR: Deep Concept Injection into Unified Autoregressive Models for Personalized Text-to-Image Generation

Supplementary Material

1. Training and Evaluation Details

Training Details. The training details of DCoAR on different datasets are shown in Table 1. For the style personalization task, only the provided single reference image is utilized. Consequently, the hyperparameters for Dual Prior Preservation (DPP) are not applied.

Evaluation Details. For the subject-style generation task, we selected 8 subjects from DreamBench [8], including 4 animals and 4 objects, as well as 6 styles from StyleDrop [9]. These yielded a total of 48 unique subject-style combinations, on which we evaluated our DCoAR, ZipLoRA [**?**], and B-LoRA [3]. For each combination, we assigned 10 distinct recontextualization prompts for both objects and living subjects, as illustrated in Table 2.

	DreamBench	StyleDrop
lr	1e-2	1e-2
batch size	1	1
resolution	768	768
λ_1	0.5	N/A
λ_2	0.5	N/A
α	1e-2	N/A
β	5e-4	5e-4
Training Steps	1000	600
Context token layers	9	3

Table 1. Training settings of DCoAR on DreamBench[8] and StyleDrop [9] datasets.

2. Details of Parameter Calculation

For PersonalAR [10], we directly reference the parameter count reported in its original paper. As for Proxy-Tuning [12], since the paper does not provide explicit parameter statistics and the code is not publicly available, we provide a coarse estimate based on the hyperparameter configurations described in the paper. Proxy-Tuning is a two-stage approach where a diffusion model is first trained, followed by an autoregressive (AR) model that is trained on data generated by the diffusion model. Both models are fine-tuned using Low-Rank Adaptation (LoRA) [4]. In our estimation, the diffusion model is based on Stable Diffusion 3.5 (SD3.5 Large) [2], while the AR model is Lumina-mGPT 7B [5]. We estimate the total number of trainable LoRA parameters based on the typical LoRA application settings.

2.1. LoRA Parameter Calculation

For a weight matrix $W \in \mathbb{R}^{d\times d}$, LoRA introduces two trainable matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$, where r is the LoRA rank. The number of trainable parameters for each projection is:

$$\text{LoRA params} = 2dr,$$

We assume LoRA is applied to all attention projections: W_q, W_k, W_v, and W_o.

Diffusion Model (SD3.5)

SD3.5 is based on the DiT [7] architecture, with 24 Transformer [11] blocks. Each block includes one self-attention and one cross-attention module. Assuming a hidden size $d = 3072$, the LoRA (rank=64) parameters per attention module are:

$$\begin{aligned} 4 \times 2 \times d \times r &= 8dr \\ &= 8 \times 3072 \times 64 \\ &= 1{,}572{,}864 \approx 1.57M. \end{aligned} \tag{1}$$

There are a total of $24 \times 2 = 48$ attention modules, leading to:

$$48 \times 1.57\text{M} = 75.5\text{M} \tag{2}$$

2.2. AR Model (Lumina-mGPT 7B)

For the AR model, we assume a hidden size $d = 4096$ and 32 Transformer layers. Applying LoRA (rank = 64) to the four attention projections in each layer results in:

$$\begin{aligned} 4 \times 2 \times d \times r = \\ &= 8dr \\ &= 8 \times 4096 \times 64 \\ &= 2{,}097{,}152 \approx 2.097M, \end{aligned} \tag{3}$$

leading to,

$$32 \times 2.097\text{M} = 67.1\text{M} \tag{4}$$

2.3. Total Parameter Count

Summing both components, the total number of trainable LoRA parameters under Proxy-Tuning is approximately:

$$75.5\text{M} + 67.1\text{M} = 142.6\text{M} \tag{5}$$

This estimate assumes LoRA is applied to all attention projections and excludes feed-forward layers. Including FFNs would increase the total parameter count further, but standard LoRA configurations typically restrict adaptation to attention modules only.

Subject	Context	Subject	Context
Objects	in the snow	Living Subjects	wearing a hat
	on the beach		with a crown
	with a city in the background		riding a bicycle
	on top of a dirt road		sleeping
	on top of green grass		in a boat
	in the jungle		in the jungle
	on the mountain		on the mountain
	floating in water		floating in water
	on a picnic table		on a picnic table
	on top of a wooden floor		playing with a ball

Table 2. Prompt templates categorized by subject type.

3. Additional Experiments

3.1. Ablation Study on Dual Prior Preservation (DPP)

To validate the design of our Dual Prior Preservation (DPP) strategy, we evaluate the individual contributions of the Class-based Next Token Prediction loss ($\mathcal{L}_{NTP_{cls}}$) and the KL Divergence constraint (D_{KL}). As shown in Table 3, applying these components in isolation leads to sub-optimal performance, highlighting the necessity of their joint application.

Specifically, employing **only** $\mathcal{L}_{NTP_{cls}}$ (Row 2) results in a significant drop in CLIP-T ($0.3192 \rightarrow 0.3011$). This indicates that a hard reconstruction constraint without distributional regularization causes the model to overfit the visual appearance of the reference, leading to severe language drift and impaired re-contextualization capabilities.

Conversely, applying **only** D_{KL} (Row 3) maintains text alignment but causes a sharp decline in subject fidelity (DINO drops to 0.7023, CLIP-I to 0.7794). This suggests that the soft distribution constraint alone is too conservative, preventing the model from learning the unique, fine-grained details of the specific subject.

However, when **combined** (Row 4), the two components exhibit a strong synergistic effect. The $\mathcal{L}_{NTP_{cls}}$ term ensures high subject fidelity (boosting CLIP-I to $\mathbf{0.8151}$), while the D_{KL} term effectively regularizes the distribution to prevent language drift, maintaining a high CLIP-T score of 0.3184. This confirms that the full DPP loss is essential for balancing identity preservation and text controllability.

3.2. Visualization of the effects of CASR.

Figure 1 provides a visual illustration of the effect of our proposed CASR loss. When the loss weight is too small, the model tends to overfit: although subject fidelity remains high, generalization to novel contexts is limited. Conversely, when the weight is too large, the context tokens become under-optimized, leading to a significant degradation in subject fidelity.

Components		Metrics		
$\mathcal{L}_{NTP_{cls}}$	D_{KL}	**DINO** (↑)	**CLIP-I** (↑)	**CLIP-T** (↑)
		0.7194	0.7905	0.3192
✓		0.7188	0.7893	0.3011
	✓	0.7023	0.7794	0.3189
✓	✓	**0.7226**	**0.8151**	0.3184

Table 3. Ablation study on the individual components of the Dual Prior Preservation (DPP) loss using the DreamBench dataset.

Figure 1. Visualization of CASR.

3.3. Joint Analysis of Modality and Insertion Depth

Figure 2 provides a comprehensive visualization of how different token modalities behave across varying insertion depths (injecting tokens into the first N layers, where $N \in \{1, 3, 9, 24\}$).

Impact of Modalities (Rows).

- **Only Text Tokens ($p_{[v]}$):** As seen in the second row, relying solely on text tokens results in *poor identity preservation* across all depths. Although the text alignment is high (e.g., the "oil painting" style in Col 4 is correctly rendered), the generated backpacks are generic and lack the fine-grained details of the reference subject.
- **Only Image Tokens (p_I):** The first row shows that while image tokens preserve visual details, they suffer from *rigidity and poor instruction following*, especially at deeper insertion layers.
- **Image and Text Tokens (Ours):** The third row demonstrates that combining both modalities ensures robust iden-

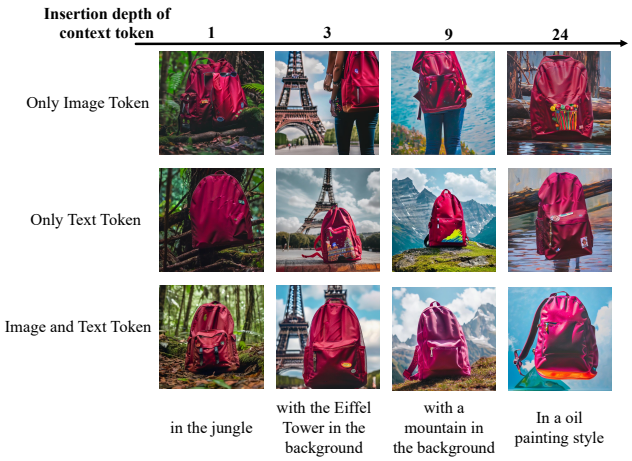

Figure 2. **Qualitative ablation of token modalities across different insertion depths.** Rows represent different token configurations, while columns correspond to the number of Transformer layers (1, 3, 9, 24) into which tokens are injected. **Rows:** "Only Text Token" suffers from identity loss (generic backpacks), while "Only Image Token" exhibits high fidelity but poor editability (e.g., failing to stylize in Col 4). **Columns:** Inserting tokens into all layers (Depth 24) leads to overfitting, preventing style changes, whereas the optimal depth (Depth 9) achieves the best balance between subject fidelity and textual control.

tity preservation while remaining responsive to text prompts.

Impact of Insertion Depth (Columns).

- **Shallow Injection (Depth 1–3):** In the early columns, the subject identity is not fully consolidated. For example, at Depth 1, the backpack's texture and shape appear slightly inconsistent with the reference, indicating insufficient visual signal propagation.
- **Deep Injection (Depth 24):** The last column (Depth 24) reveals the detrimental effect of *over-injection*. In the "Only Image Token" setting, the backpack fails to transform into the "oil painting" style and remains photorealistic. This indicates that injecting visual tokens into all layers creates an overly strong visual prior that overrides the textual style control, leading to **overfitting**.
- **Optimal Depth (Depth 9):** Our chosen setting (Depth 9, Column 3) achieves the optimal sweet spot. The model successfully retains the specific identity of the backpack (unlike the text-only baseline) while seamlessly integrating it into the new background (unlike the depth-24 baseline), validating our decision to inject concepts only into the early-to-mid layers.

3.4. Additional Qualitative Results

Figures 4 to 7 presents additional qualitative results of subject-driven personalization and style personalization.

We provide further visual evidence comparing DCoAR with competing baselines in Figure 3. Consistent with our

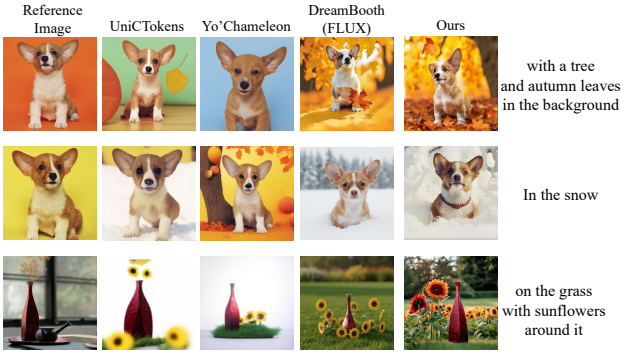

Figure 3. **Additional qualitative comparison of subject-driven personalization**. We compare our DCoAR against two representative concept-injection methods (UniCTokens [1], Yo'Chameleon [6]) and a state-of-the-art diffusion-based fine-tuning approach (DreamBooth [8] with FLUX).

main paper's findings, shallow concept-injection methods (UniCTokens[1], Yo'Chameleon[6]) struggle to capture high-frequency details, often resulting in smoothed textures and lower visual fidelity. Conversely, while the adaptation-based method DreamBooth (utilizing the FLUX backbone) produces high-quality backgrounds, it exhibits significant semantic instability and identity drift; this is particularly evident in the second row, where the reference dog is erroneously rendered as a different breed. In contrast, DCoAR successfully preserves intricate subject details—such as specific fur patterns and material textures—while accurately adhering to complex background prompts, demonstrating the effectiveness of our Deep Concept Injection strategy.

Figures 5 and 6 demonstrate that given only a few reference images, DCoAR successfully performs subject-driven generation across various tasks, including recontextualization (e.g., novel backgrounds), property modification (e.g., color and shape), and accessorization (e.g., adding glasses or outfits). The results demonstrate strong subject fidelity, adaptability to novel contexts, and fine-grained controllability.

Figure 4 illustrates the effects of diverse contextual prompts on image generation for various subjects across different visual styles. Each row corresponds to a distinct subject, while each column varies the context.

4. Limitations

Despite the superior performance of DCoAR in personalized generation, we acknowledge the primary limitations that warrant future investigation: **Sensitivity to Insertion Depth and Overfitting.** Our framework relies on a carefully tuned Layer-wise Multimodal Context Learning (LMCL) strategy. As evidenced by our ablation studies, the model is sensitive to the depth of token injection. While shallow injection fails to capture identity details, **injecting to-**

Figure 4. Impact of contextual prompts on image generation across subjects and styles.

kens into overly deep layers imposes excessive visual constraints. This leads to *overfitting*, where the model becomes rigid and prioritizes pixel-level adherence to the reference image over the semantic control of the text prompt, hindering effective re-contextualization and stylization. Developing an adaptive mechanism to automatically determine the optimal insertion depth for different subjects remains an open problem.

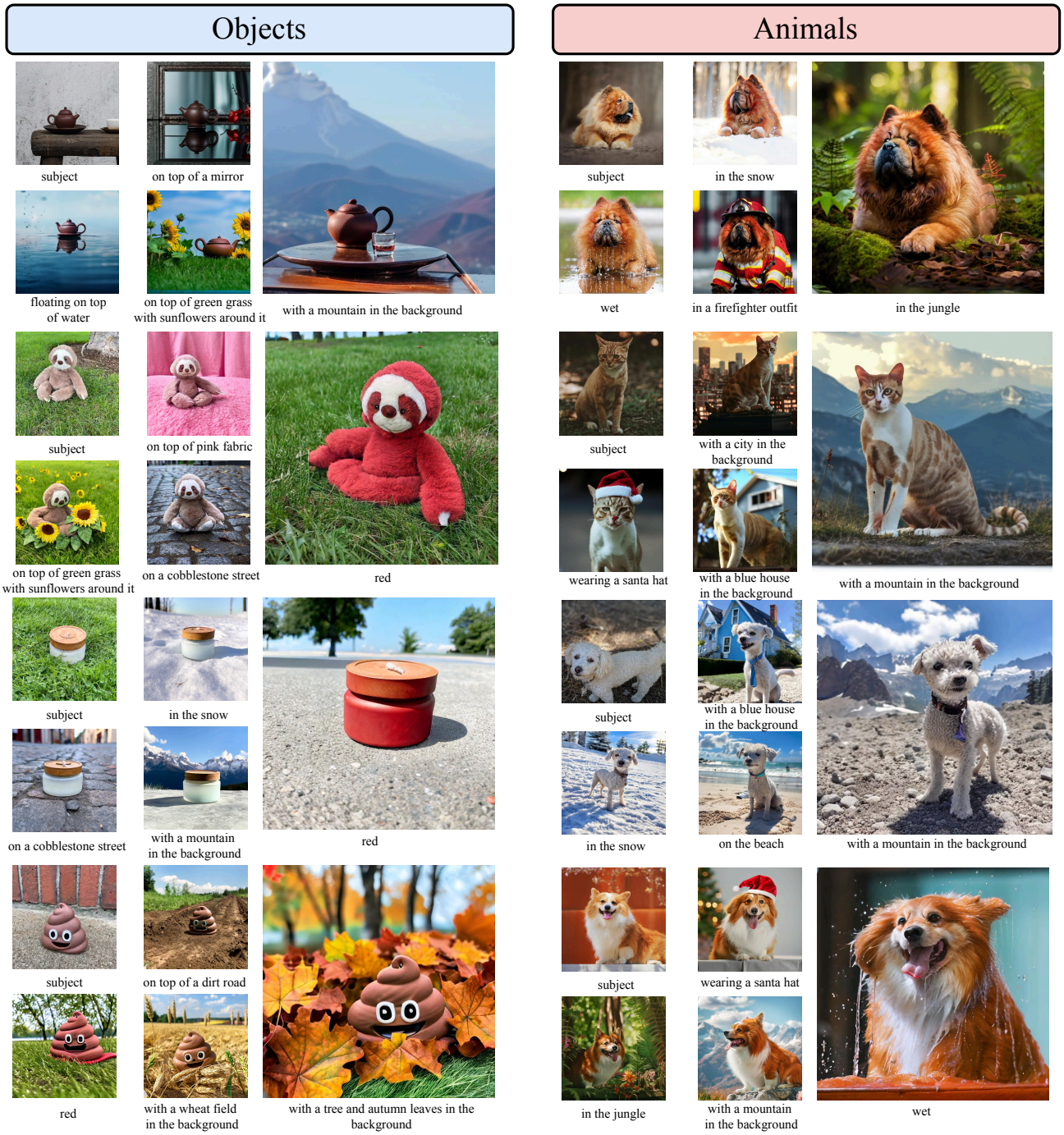

Figure 5. Qualitative results of subject-driven customization with DCoAR.

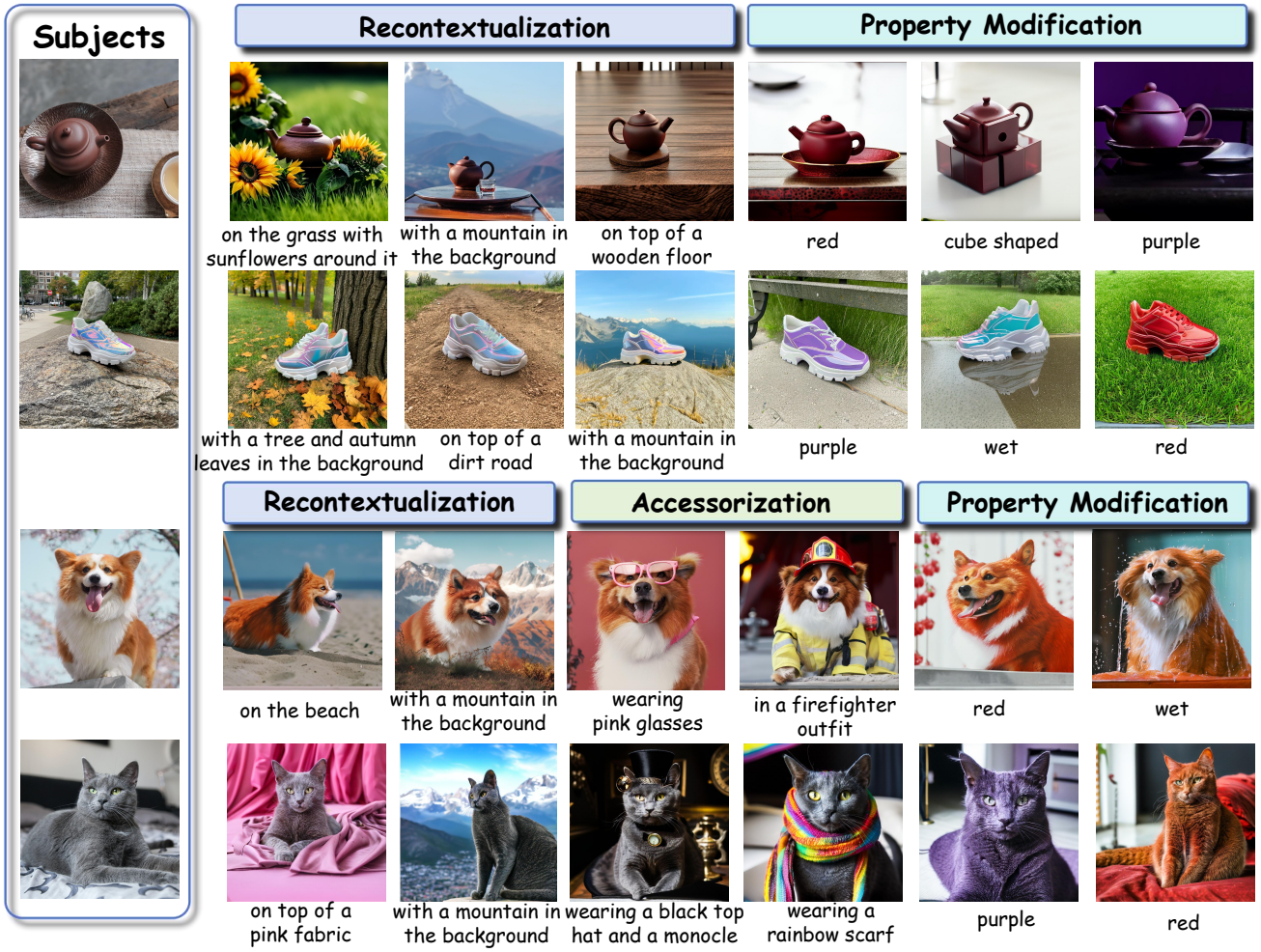

Figure 6. Qualitative results of subject-driven customization with DCoAR.

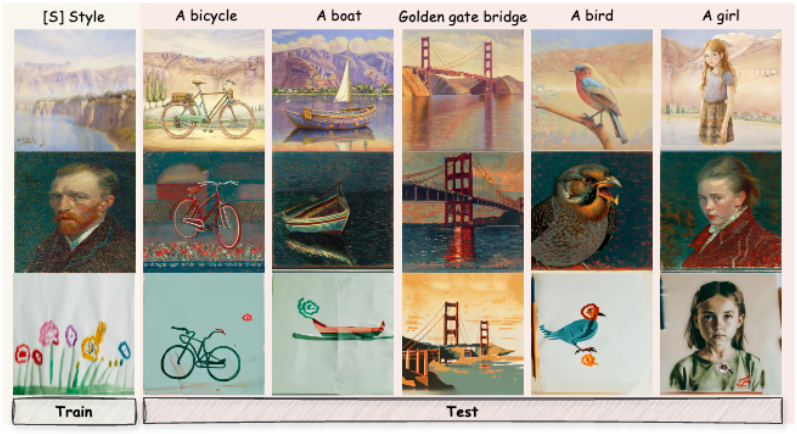

Figure 7. DCoAR demonstrates effective style personalization conditioned on only a single reference image.

References

[1] Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, Bocheng Zou, Chaoqun Yang, and Wentao Zhang. Unictokens: Boosting personalized understanding and generation via unified concept tokens, 2025.

[2] 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 *ICML*, 2024.

[3] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *ECCV*, pages 181–198, 2024.

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[5] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.

[6] Thao Nguyen, Krishna Kumar Singh, Jing Shi, Trung Bui, Yong Jae Lee, and Yuheng Li. Yo'chameleon: Personalized vision and language generation. In *CVPR*, pages 14438–14448, 2025.

[7] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.

[8] 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 *CVPR*, pages 22500–22510, 2023.

[9] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

[10] Kaiyue Sun, Xian Liu, Yao Teng, and Xihui Liu. Personalized text-to-image generation with auto-regressive models. *arXiv preprint arXiv:2504.13162*, 2025.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[12] Yi Wu, Lingting Zhu, Lei Liu, Wandi Qiao, Ziqiang Li, Lequan Yu, and Bin Li. Proxy-tuning: Tailoring multimodal autoregressive models for subject-driven image generation. *arXiv preprint arXiv:2503.10125*, 2025.