Appendix

A. Implementation Details

A2. Architecture

We leverage the advanced pre-existing knowledge from pretrained MLLMs to aid in model convergence. Specifically, we initialize our multi-modal decoder with LLaVA-v1.5-7B checkpoint [2] from HuggingFace. Its vision encoder is CLIP-vit-large-patch14 (CLIP-L) [5] and takes in I^{wb}_{ref}. We use ViT-H/1 [3] to extract the image embedding of I_{ref}. Thus, a trained linear mapping is added at the output of MLLM to map the learned text token from $\mathbb{R}^{4096}$ to $\mathbb{R}^{1024}$. The model is built on top of the IP-Adapter [7] and the training code on LDM [6]. Our foundation model is StableDiffusion v1.5 [6] with Realistic-Vision-v4.0 [1] checkpoint. Both subject and context cross-attentions are implemented with the diffusers attention processor class. We call our MLLM as multi-modal decoder because: it takes in visual features of I^{wb}_{ref} and text instruction and "generates the full image I_{ref}" in the form of CLIP embeddings. So our MLLM is not a typical subject-feature encoder like in most existing works, but instead a generative MLLM decoder.

A2. Training

we lock the parameters of the image encoders and only adjust the MLLM parameters along with the newly incorporated learnable token. In stage two training, the image/text encoder, multi-modal decoder, and the entirety of the UNet parameters are frozen. We optimize exclusively the linear mappings of subjet-cross-attention and context-cross-attentions.To expedite convergence, we employ attention mapping layers from the IP-adapter [7] and start with a zero-initialized subject-cross-attention. We apply large gradient accumulation steps (200) and clip grad norm for smoother training. In both stages, we use the AdamW optimizer with a learning rate of 1^{-5}. Training takes 7 days in total on 16 A100 80G GPUs.

A3. Evaluation

We use the DDIM sampler with default parameters and 50 denoising steps. Evaluation requires 23.5G GPU RAM for 512×512 resolution. The iterative masking is disabled in the first denoising step since $M_{gen(t)}$ is unavailable yet. We apply classifier-free guidance on both the text prompt and Context-cross-attention side, but not the subject-cross-attention. On the text prompt side, simply replace the target prompt P_{tgt} with null embedding. On the context-cross-attention side, we zero out the contextualized features before feeding them to attention linear mappings. We use $\beta = 1.0$ in the recontextualization task for better subject detail accuracy and, empirically, $\beta = 0.5$ in the texture editing task for better balance between subject and prompt.

B. Visual comparison and more Results

B1. Uncurated Comparison with Baselines

We show uncurated results for our model and the baselines. Results are random and not selected. For each method, 8 images are generated. IP-Adapter failed to accurately capture the structure of the car in Fig. 1. In Fig. 2, increasing the weight/scale to 0.85 in IP-Adapter still yields a wrong color pattern of the dog. Under this scale, The original white background strongly interferes with its generated image. Ours correctly captures the color and structure pattern with much more favorable backgrounds, respecting both the reference image and the text prompt.

We conclude from above that:(i) Our multi-modal LLM decoder and its contextualized feature are crucial to our model performance. It extracts visual features from reference image I^{wb}_{ref} and contextualizes it with the target text prompt P_{tgt}. Its output is trained to match the overall CLIP image embedding of the resulting image and provides critical features to bridge the gap between the subject feature and the target prompt. (ii) Iterative masking helps distinguish between the background and main subject, improving CLIP-T. Removing it leads to corrupted backgrounds and significantly lower image quality. (iii) Disabling the Subject-cross-attention worsens the CLIP-I score as features from MLLM are insufficient to accurately reproduce subject details. This quantitative ablation lines up with our intuition: our multi-modal decoder and its context-cross-attention provide vital image features and serve as a foundation. Subject-cross-attention and iterative masking help in subject detail accuracy and background quality respectively.

B3. More Results

In Tab. 1, we show detailed quantitative results for the DreamBooth dataset. From Fig. 3 to Fig. 8, we provide additional qualitative results on various subjects and prompts. Reference subject images are on the left. In the rest columns, we provide generated renditions. Fig. 3 to Fig. 6 are about context editing and Fig. 7 and Fig. 8 are about texture editing.

C. Discussions

C1. Social Impact

While tuning-based personalization models are largely inaccessible to most people because of computation resource limits, our method of open-vocab tuning-free personalization model helps democratize such models to everyday users with a significantly improved quality. However, it also bears the potential risk of being exploited for the creation of deceptive content or the propagation of misinformation. To address this concern, we have specifically designed our training process to exclude person-related subjects and focus on generic objects. This intentional

	Ours				IP-adapter		Blip-diffusion	
	DINO	CLIP-I	CLIP-$T_{(c)}$	CLIP-$T_{(t)}$	CLIP-$T_{(c)}$	CLIP-$T_{(t)}$	CLIP-$T_{(c)}$	CLIP-$T_{(t)}$
backpack	0.590	0.861	0.350	0.338	0.342	0.337	0.324	0.292
backpack2	0.438	0.698	0.361	0.351	0.357	0.365	0.323	0.296
boot	0.502	0.819	0.346	0.345	0.330	0.325	0.287	0.277
bowl	0.621	0.699	0.350	0.338	0.300	0.339	0.249	0.257
can	0.657	0.749	0.361	0.321	0.355	0.347	0.310	0.294
candle	0.474	0.734	0.361	0.353	0.348	0.330	0.329	0.297
cartoon	0.574	0.794	0.305	0.312	0.284	0.278	0.255	0.248
cat	0.813	0.840	0.354	0.339	0.330	0.300	0.320	0.261
cat2	0.750	0.845	0.358	0.345	0.327	0.310	0.292	0.263
clock	0.679	0.866	0.335	0.341	0.305	0.324	0.322	0.270
dog	0.725	0.863	0.342	0.334	0.328	0.308	0.302	0.261
dog2	0.652	0.863	0.338	0.335	0.322	0.301	0.302	0.260
dog3	0.609	0.799	0.342	0.345	0.326	0.304	0.303	0.257
dog5	0.574	0.761	0.344	0.340	0.298	0.279	0.324	0.257
dog6	0.713	0.857	0.332	0.324	0.320	0.300	0.282	0.253
dog7	0.709	0.850	0.338	0.338	0.325	0.287	0.305	0.248
dog8	0.654	0.842	0.337	0.333	0.316	0.289	0.306	0.246
plushie	0.620	0.781	0.367	0.334	0.341	0.315	0.304	0.287
plushie2	0.491	0.754	0.372	0.333	0.340	0.336	0.326	0.299
plushie3	0.621	0.794	0.366	0.323	0.343	0.317	0.289	0.271
poop	0.533	0.732	0.345	0.339	0.335	0.338	0.327	0.291
sneaker	0.665	0.814	0.350	0.347	0.333	0.312	0.276	0.256
sneaker2	0.704	0.824	0.340	0.346	0.325	0.305	0.269	0.264
sunglasses	0.655	0.851	0.349	0.344	0.339	0.348	0.299	0.293
teapot	0.642	0.852	0.383	0.361	0.366	0.359	0.294	0.288
toy1	0.508	0.765	0.341	0.306	0.332	0.292	0.292	0.258
toy2	0.627	0.821	0.334	0.310	0.320	0.328	0.287	0.260
toy3	0.532	0.741	0.348	0.313	0.337	0.311	0.292	0.266
vase	0.593	0.815	0.358	0.339	0.338	0.356	0.295	0.299
Average	0.618	0.803	0.348	0.335	0.330	0.319	0.300	0.271

Table 1: Quantitative details on DreamBooth datasets. We add 10 new prompts focusing on texture editing to calculate CLIP-$T_{(t)}$. These prompts are: A sculpture of a *label* made of (Lego/Paper/Gold/Wood/silver/green jade/glass/stone/sketch/Minecraft).

Fig. 1: Uncurated random renderings from our model and baselines. IP-Adapter failed in capturing car structures, despite increasing its scale from default 0.6 to 0.85. Our results have better detail accuracy, more diverse composition, and favorable quality than the baselines.

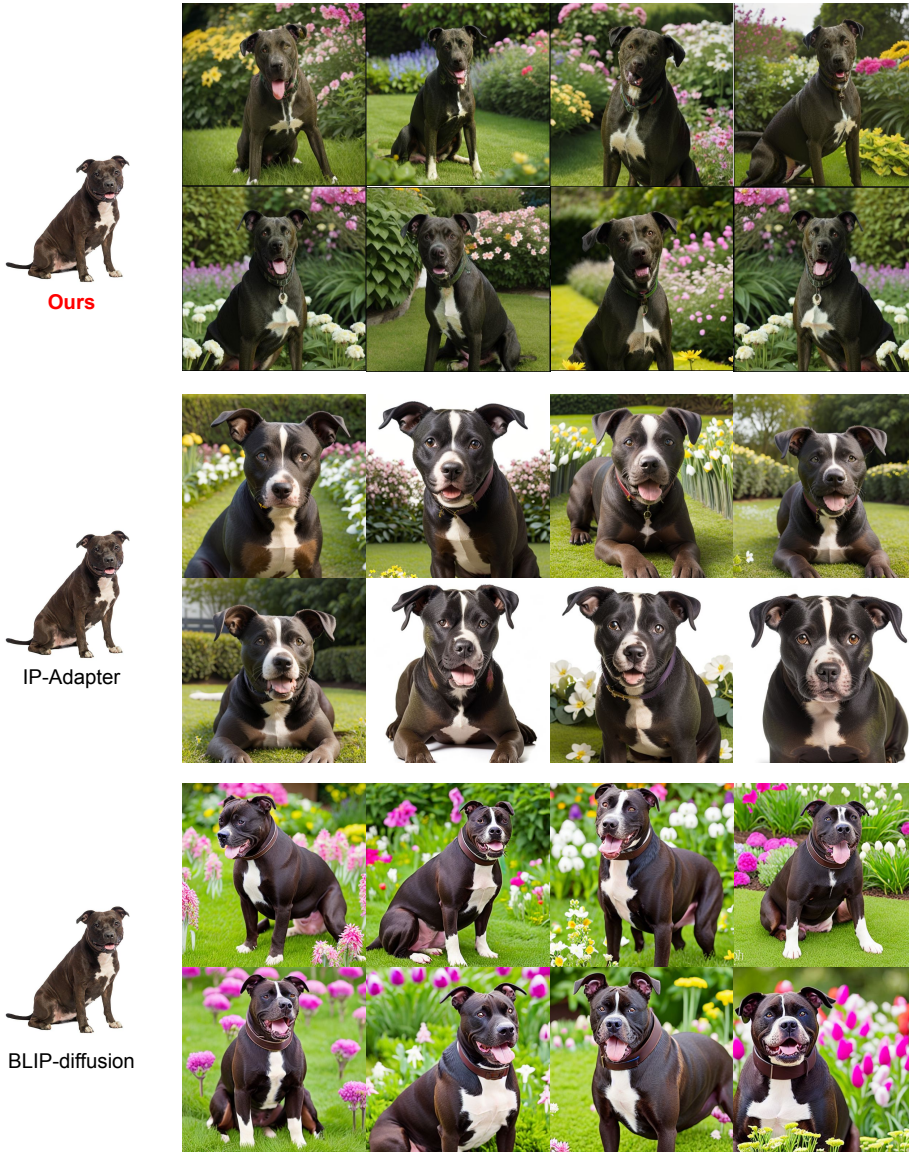


Fig. 2: Uncurated random renderings from our model and baselines. IP-adapter generates inaccurate color patterns, even with an increased scale(weight) to 0.85. Since it does not distinguish background from foreground, some results failed to follow the "garden with flower" in the prompt. Our results have more accurate details with much better backgrounds.

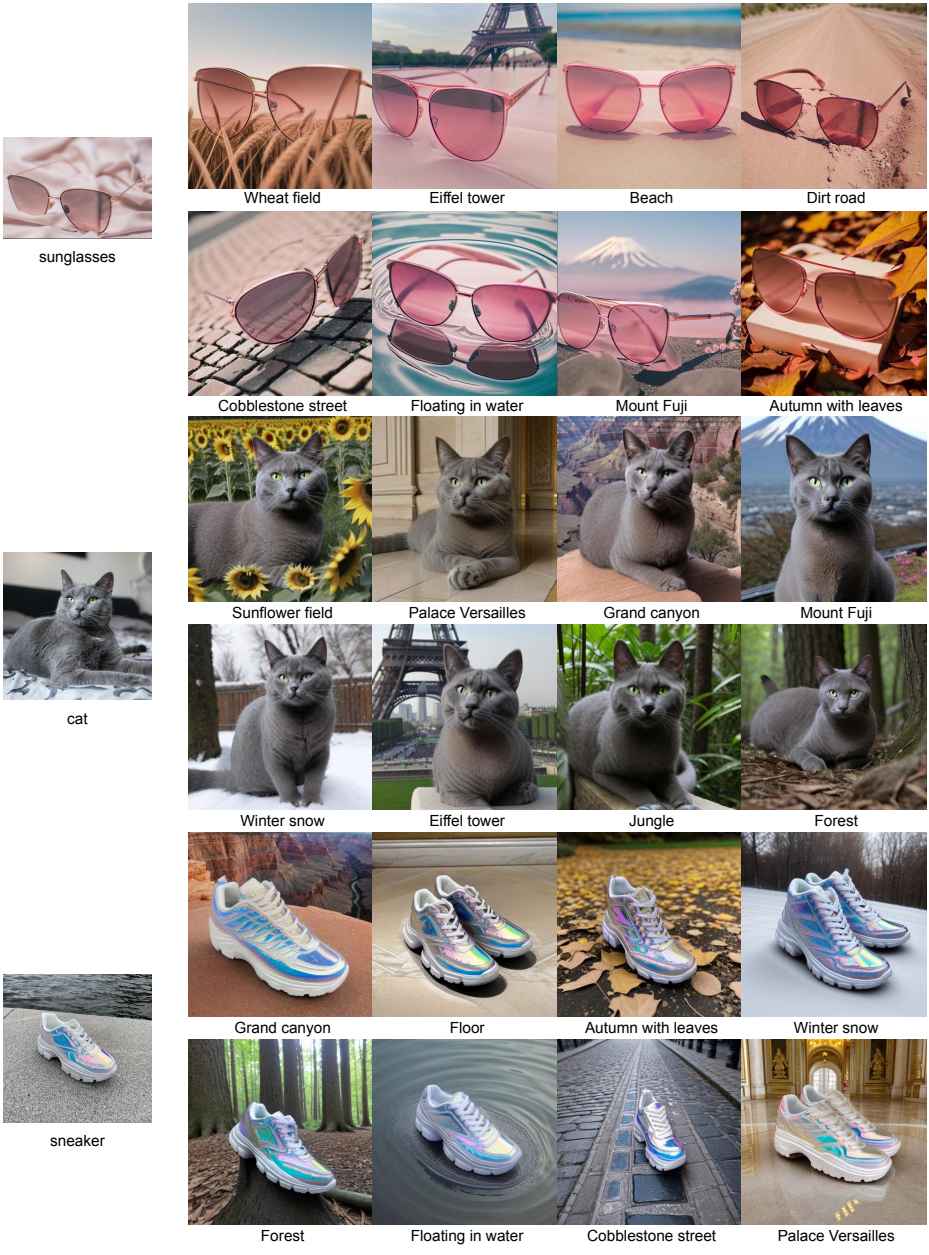


Fig. 3: Additional results from our model. Change context.

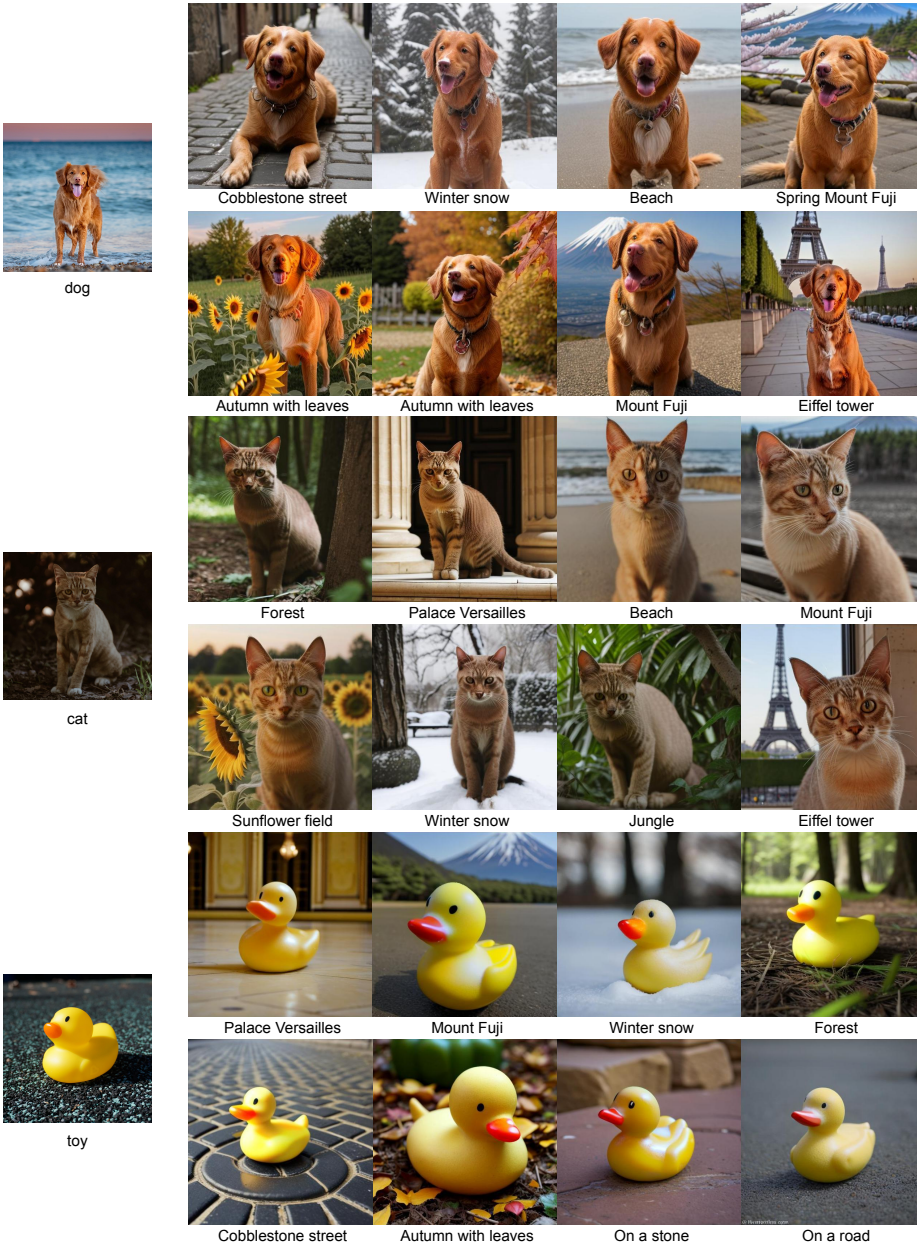


Fig. 4: Additional results from our model. Change context.

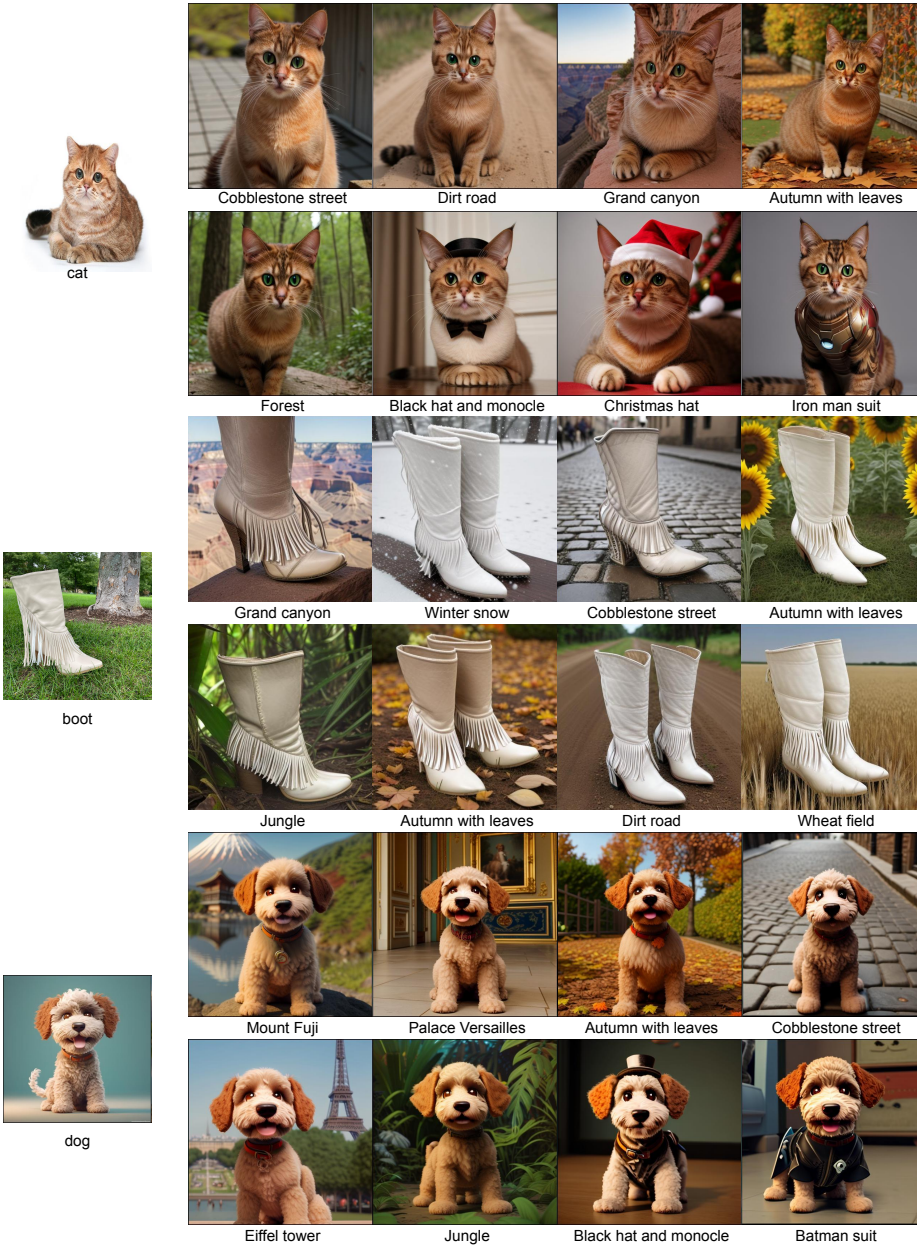


Fig. 5: Additional results from our model. Change context.

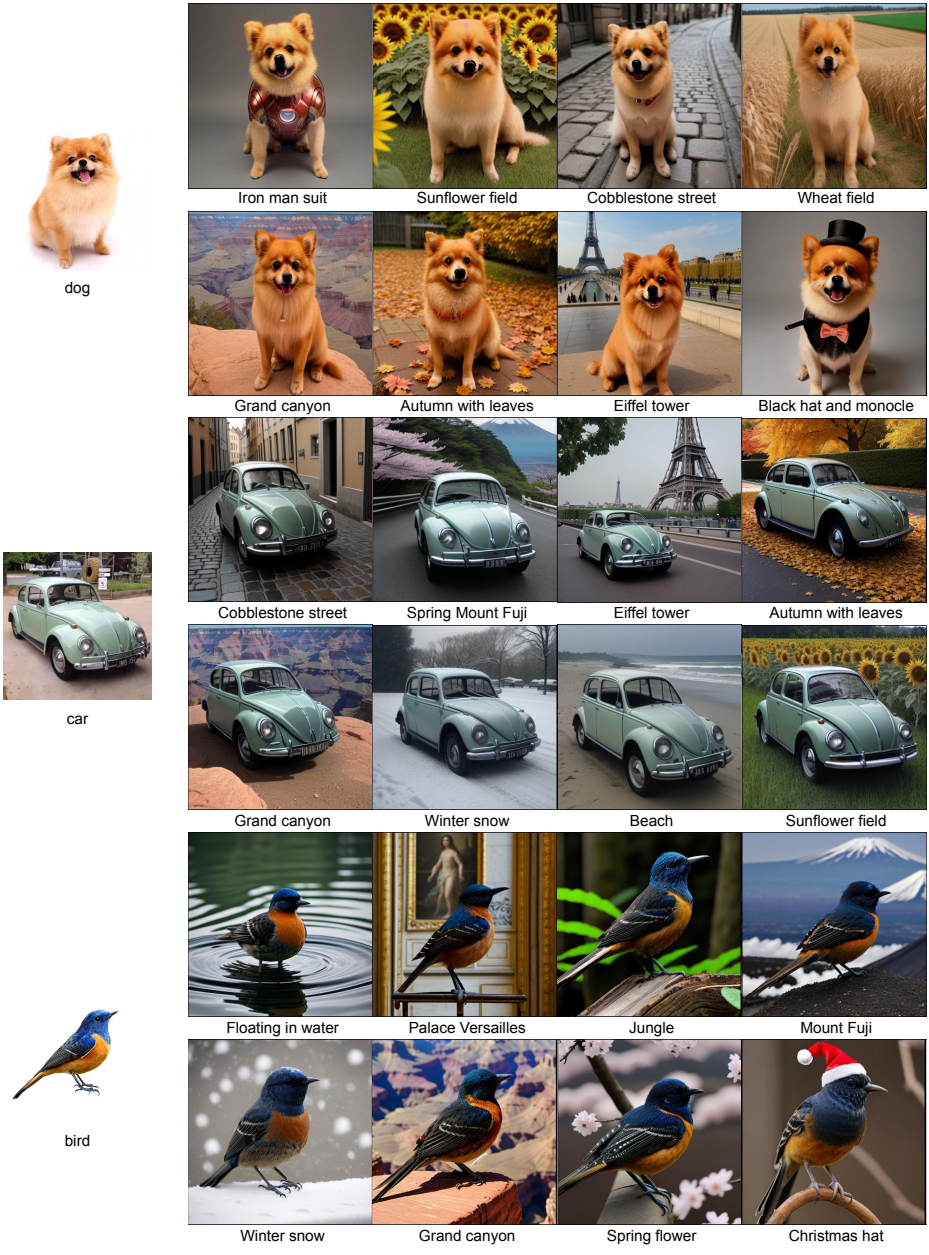


Fig. 6: Additional results from our model. Change context.

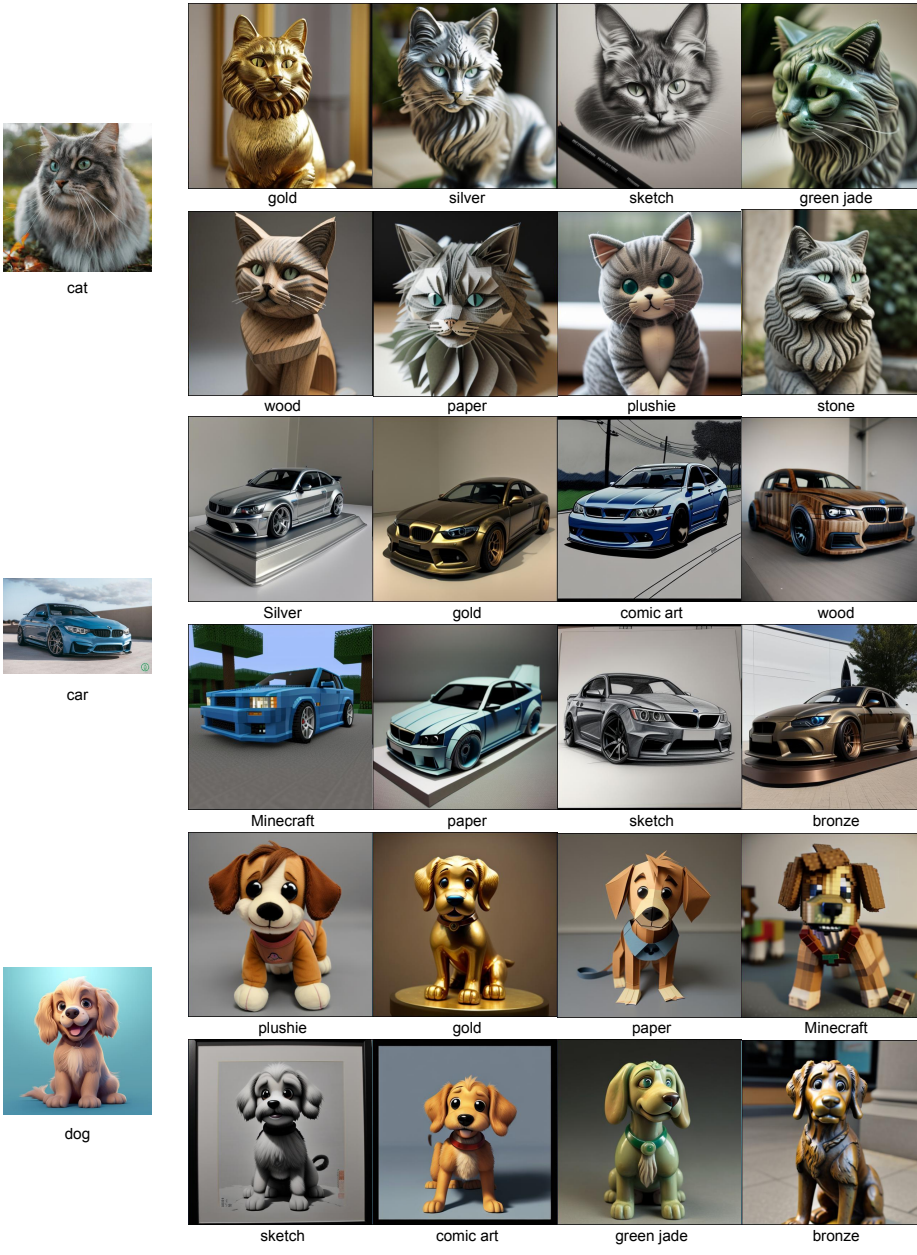


Fig. 7: Additional results from our model. Change texture.

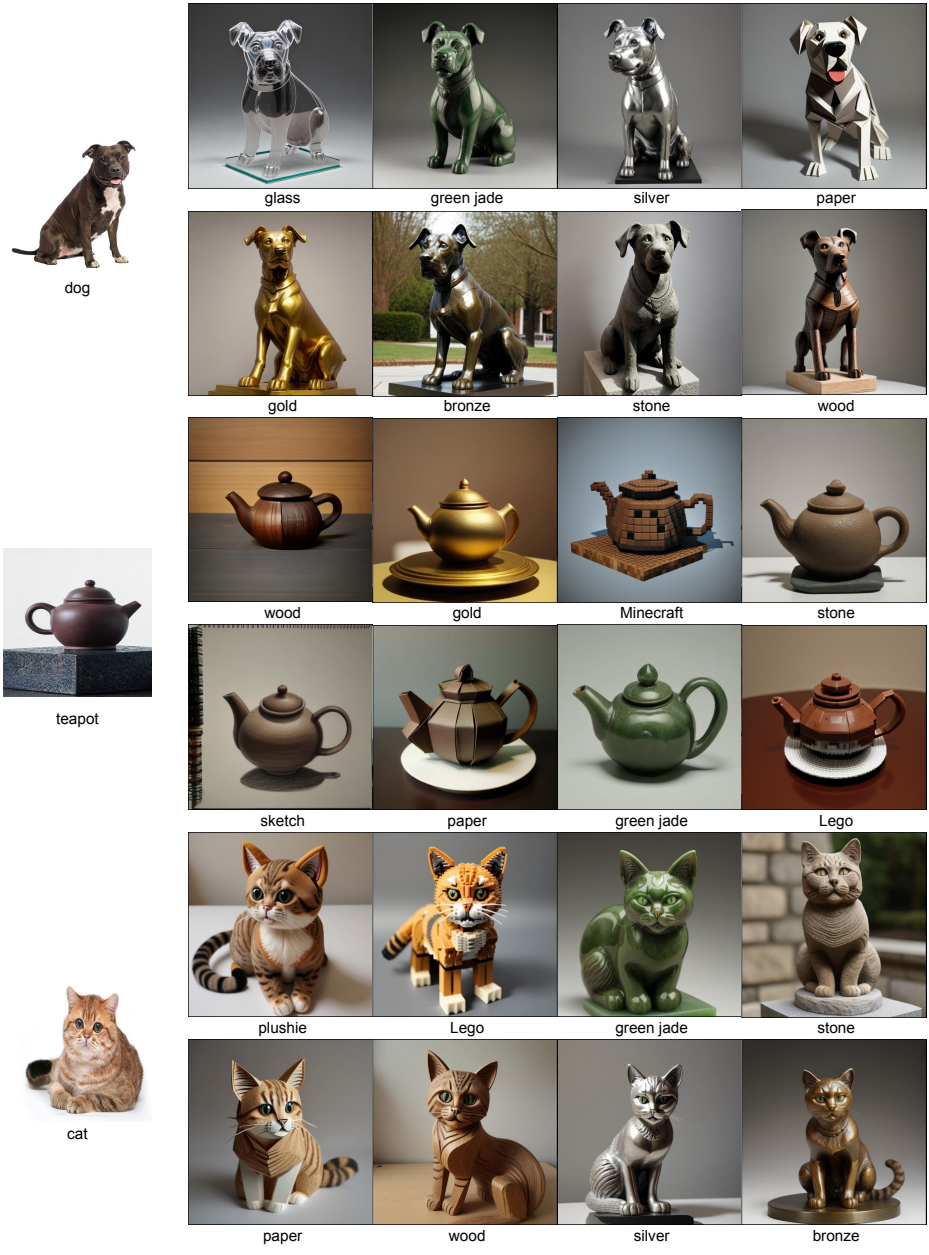


Fig. 8: Additional results from our model. Change texture.

limitation reduces the model's ability to generate convincing counterfeit images where individuals are central elements. To ensure the integrity of content generated by our model, we advise a thorough examination of its outputs before deploying our model in consumer-facing applications.

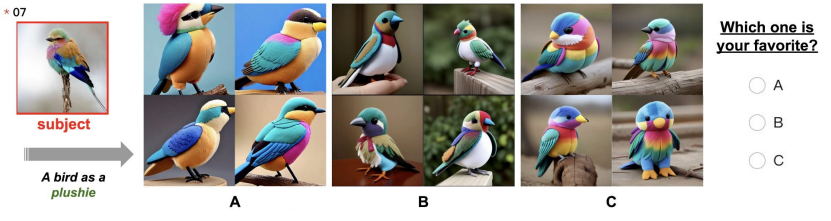
C2. Failure Examples

Fig. 9: Failure cases. The **text** messages displayed on the bowl and can appear distorted or missing in the resulting images, a flaw inherited from SD v1.5. This version is notably challenged in its ability to reproduce text with accuracy.

We also noticed that some subjects are much easier to learn than others [4]. For example, the model generates high-quality results for dogs and cats with consistent identity and almost identical details. Our improvement against baselines [4] [7] starts to be more noticeable for rare subjects like shoes and robots. Occasionally, as shown in Fig. 9, with subjects that are rarer especially accompanied by text, the model is unable to fully capture its details.

C3. About User Study

We design 9 questions: 6 for recontextualization task (3 about subject fidelity, 3 about background-prompt fidelity) and 3 for texture editing task. Nine ratings per user and a total of 774 ratings were collected. A sample is shown below:



References

1. Adhik Joshi: realistic vision v40 (2024), https://huggingface.co/stablediffusionapi/realistic-vision-v40, accessed: 2024-03-06 1
2. Haotian Liu: Llava-1.5-7b (2024), https://huggingface.co/liuhaotian/llava-v1.5-7b, accessed: 2024-03-14 1
3. Humza Iqbal: Clip-h (2024), https://github.com/mlfoundations/open_clip, accessed: 2024-03-14 1
4. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36** (2024) 12
5. OpenAI: Clip-l (2024), https://huggingface.co/openai/clip-vit-large-patch14, accessed: 2024-03-14 1
6. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1
7. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023) 1, 12