UniVerse: A Unified Modulation Framework for Segmentation-Free, Disentangled Multi-Concept Personalization

Supplementary Material

1. Data Curation

1.1. Initial Dataset and Motivation

Due to the lack of a publicly available dataset specifically designed for multi-subject personalized image generation, we created a new training resource for our UniVerse model. We initiate our pipeline using the public UNO dataset [**?**], which originally contains approximately one million image pairs, each consisting of a single-subject reference image and a target image generated by FLUX1.1dev [43]. After filtering for high-quality pairs, we collect around 300K image pairs to serve as the foundation for our training data.

1.2. Multi-Subject Data Curation Pipeline

Inspired by UNO [**?**] and XVerse [**?**], we perform a reversed data curation pipeline to convert the single-subject foundation into a multi-subject reference dataset, as illustrated in Fig. 2. This pipeline is comprehensive and designed to obtain multi-subject reference images for each target image and its corresponding prompt.

Entity Extraction and Grounding. We first extract relevant entities from the prompt using the SpaCy [42] parser, focusing on noun phrases. The OwlV2 [44] model is then used to ground these phrases by predicting bounding boxes within the target image. Abstract phrases associated with low-confidence bounding boxes are filtered out to maintain data quality in the reference set.

Segmentation and Generation. With the filtered bounding boxes and phrases, we utilize SAM2 [45] to segment the target objects. This yields the white-background reference images required by the XVerse model [**?**]. The input prompt for XVerse is generated by Qwen3 [46].

Reference Diversity. We enhance the diversity of the final dataset by combining multiple reference images for a single target image. Specifically, 75% of the generated reference images use a single subject, while the remaining 25% are created with two subjects. The output images of XVerse serve as the final reference images for the corresponding target images.

Reference Prompt Generation. To obtain accurate reference prompts describing the generated output, we use the Qwen3VL vision-language model [41].

Ultimately, this process yields a dataset with multiple-subject reference images for each target image, substantially increasing the diversity and quantity necessary for training UniVerse in complex personalization scenarios.

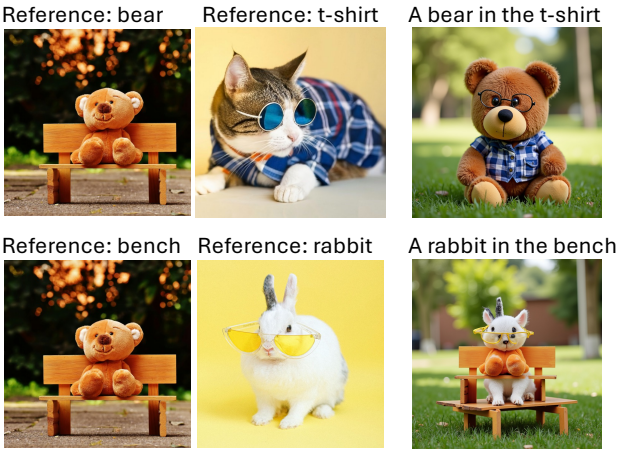

Figure 1. **Failure Cases of UniVerse in Multi-reference Images.** Here, the glasses always appear in the generated images and there is identity mixing between references.

2. Additional Qualitative Results

We include more qualitative results across diverse contexts and styles on Dreambench. With given only single-object inputs, our method consistently produces rich and high-quality results while preserving object identity across different contexts and styles. We also provide the `index.html` with more results on our benchmark, model's application and ablation studies.

3. Limitation and Future work

3.1. Additional quantitative result

In Table 1, UniVerse achieves competitive performance, showing results that are largely consistent with other strong baselines in single-subject generation.

3.2. Limitations

Our proposed approach, while effective, still presents several limitations:

Benchmark Insufficiency. The field currently lacks a comprehensive, segmentation-free benchmark tailored for multi-reference generation. The absence of a more extensive benchmark—particularly one featuring richer reference sets (*e.g.*, three or more distinct concepts, each with multiple attributes), hinders rigorous, standardized evaluation and comparison.

Concept Interference. The model is not entirely resilient

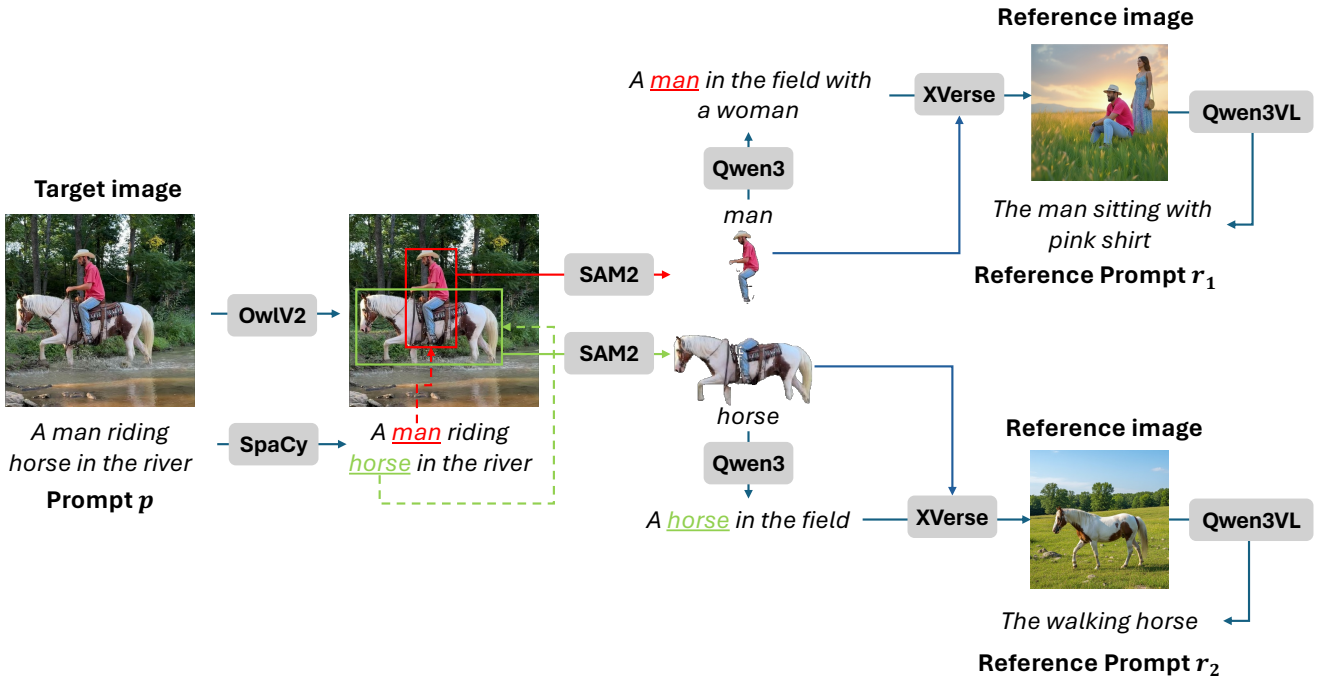

Figure 2. **Data Curation for Cross-Reference Images.** From the image with many objects, we extract each entity and adopt XVerse to generate reference images in different context.

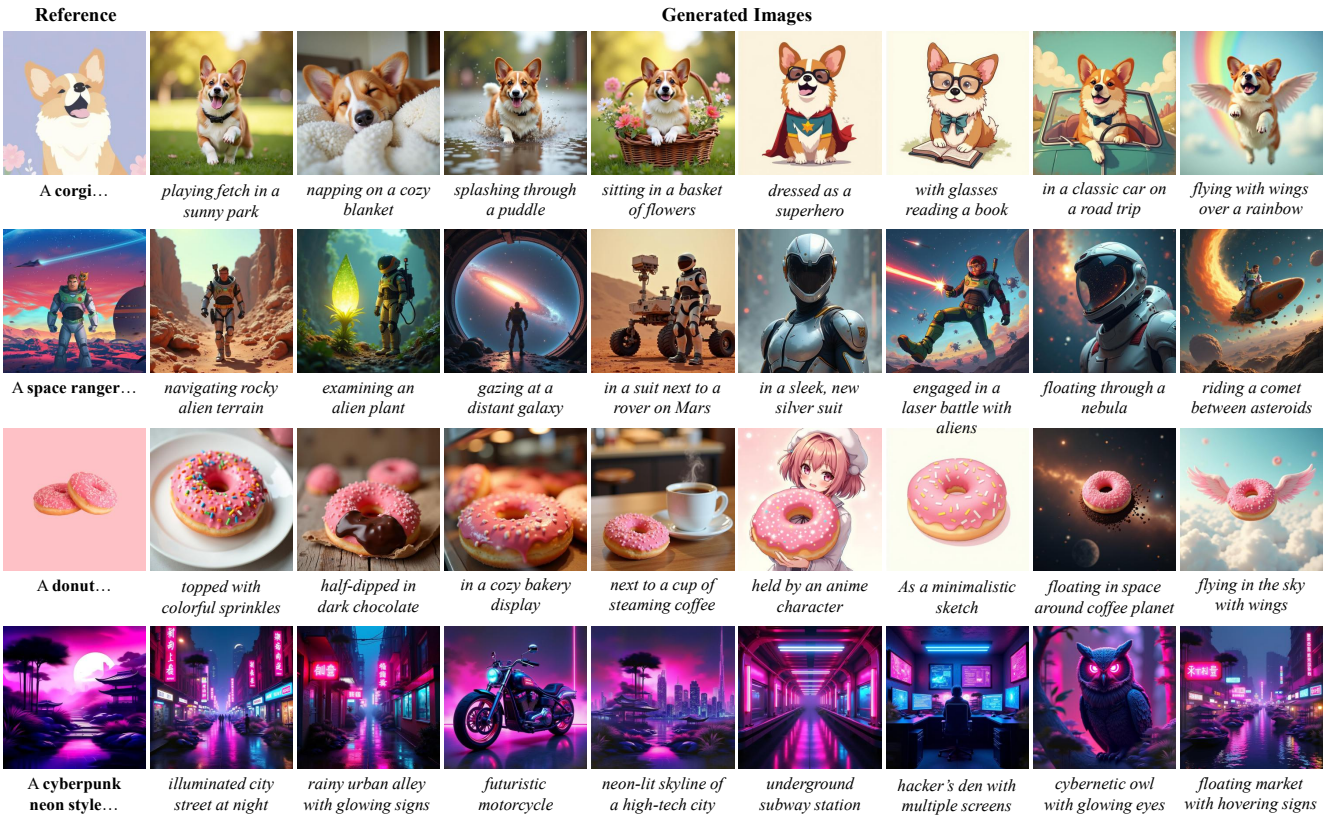

Figure 3. **Additional Qualitative Results of our models with diverse contexts and styles in Dreambench.** The generated images are in high-quality with identity preserved.

Table 1. Quantitative comparison of general single-subject personalization generation performance on DreamBench++. **Bold** represents best performance, underline represents second best.

Method	CP · PF	Concept Preservation					Prompt Following			
		Animal↑	*Human*↑	*Object*↑	*Style*↑	Overall↑	*Realistic*↑	*Style*↑	*Imaginative*↑	Overall↑
UNO [**?**]	0.514	0.650	0.435	0.653	0.325	0.579	0.947	0.869	0.792	0.887
DreamO [**?**]	0.542	0.679	0.586	0.621	0.375	0.601	0.935	0.910	0.821	0.901
OmniGen [**?**]	0.499	0.619	0.486	0.561	0.392	0.546	0.949	0.938	0.807	0.914
OmniGen2 [**?**]	0.536	0.574	0.515	0.621	0.481	0.574	**0.953**	**0.948**	**0.874**	**0.934**
MS-Diffusion [**?**]	**0.570**	**0.803**	**0.707**	**0.688**	0.614	**0.715**	0.868	0.819	0.623	0.797
MIP-Adapter [**?**]	0.375	0.670	0.599	0.538	**0.722**	0.610	0.714	0.599	0.436	0.614
XVerse [**?**]	0.514	0.687	0.668	0.568	0.536	0.613	0.902	0.802	0.761	0.838
UniVerse (Ours)	0.540	0.620	0.700	0.610	0.560	0.622	0.915	0.825	0.780	0.855

to concept interference or leakage between subjects. While using more restrictive prompts (*e.g.*, "just the cat") can partially alleviate this issue, a robust architectural solution is still necessary.

Overfitting and Prompt Sensitivity. Our method may occasionally overfit to a specific reference subject, leading to a decline in generalization. Furthermore, its performance degrades when exposed to vague or semantically incoherent prompts. Examples illustrating this limitation are presented in Fig. 1.

3.3. Future Work

There are several promising directions for extending this work and addressing the aforementioned limitations:

Benchmark Development. A critical immediate avenue is the construction of a large-scale, segmentation-free multi-reference benchmark. This will support more systematic evaluation and facilitate broader progress within the community.

Disentanglement and Adaptation. We plan to investigate developing stronger disentanglement modules or adaptive conditioning strategies to significantly reduce concept interference and improve subject separation.

Robustness via Understanding. Incorporating prompt-understanding or semantic-consistency models could substantially improve the system's robustness when processing ambiguous or low-quality textual descriptions.

Balancing Identity and Generalization. Exploring training strategies that deliberately balance identity preservation with generalization is essential to mitigate the observed overfitting to specific references.

Multimodal Extension. Finally, extending the current framework to video or 3D domains offers an exciting opportunity for exploring broader multimodal applications.

References

[41] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[42] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020. 1

[43] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1

[44] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *NeurIPS*, 2023. 1

[45] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[46] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1