Supplementary Material of
UniVG: A Generalist Diffusion Model for Unified Image Generation and Editing

Comparison with ACE++ [2] While they adopt channel-wise concatenation of latent noise, image, and mask (Eq. 4), UniVG builds on this idea to support task-unified modeling with fixed-length MM-DiT inputs across a broader task spectrum, including tasks not addressed in ACE++, such as layout-guided generation and ID customization. Unlike them, which trains multiple vertical models with task-specific finetuning, UniVG employs a single model with a unified conditioning abstraction that flexibly incorporates images, masks, text, layouts, and face embeddings. Our three-stage curriculum (Table 5&6) is empirically designed to manage task interference (*e.g.*, ID hindering editing) and to leverage synergy (*e.g.*, depth aiding editing), an aspect not explored by ACE++. In addition, our UniVG achieves a greater inference efficiency by maintaining fixed-length inputs, thereby avoiding the scaling limitations of sequence concatenation used in ACE++ (Table 7).

Comparison with OmniGen [5] and OneDiff [1] Both OneDiff and OmniGen concatenate additional images with the noise, resulting in an excessively long sequence that significantly increases computational overhead (Table 7). Furthermore, we believe such generalist design choices, especially when supported by empirical results and training insights, represent a meaningful step forward in building deployable foundation models.

Results on Text-guided Inpainting The comparisons in text-guided inpainting on EditBench [4] are shown below. UniVG even outperforms inpainting models, where we can recover the masked region that is more aligned with the given prompt (T2I) as well as the reference image (I2I).

CLIP-Score	SD	DL2	IM	SDXL	UniVG
T2I ↑	29.7	29.1	31.5	30.4	**33.6**
I2I ↑	74.9	76.1	76.6	82.5	**87.3**
Mean ↑	52.3	52.6	53.6	56.4	**60.5**

Results on Depth Estimation We follow OneDiff [1] and evaluate the depth estimation task on DIODE [3]. Our unified model surpasses prior task-specific methods, such as DivDep and MiDaS, and achieves competitive performance with LeReS. This highlights the potential of UniVG to support precise vision tasks.

Metric	DivDep	MiDaS	LeReS	DA2	OneDiff	UniVG
AbsRel ↓	37.6	33.2	**27.1**	<u>27.1</u>	29.4	32.2
δ_1 ↑	63.1	71.5	**76.6**	74.8	75.2	<u>75.3</u>

References

[1] Duong Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One Diffusion to Generate Them All. In *arXiv:2411.16318*, 2024. 1

[2] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. ACE++: Instruction-Based Image Creation and Editing via Context-Aware Content Filling. In *arXiv:2501.02487*, 2025. 1

[3] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. In *arXiv:1908.00463*, 2019. 1

[4] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[5] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified Image Generation. In *arXiv:2409.11340*, 2024. 1