HBridge: H-Shape Bridging of Heterogeneous Experts for Unified Multimodal Understanding and Generation

Supplementary Material

Xiang Wang[1] Zhifei Zhang[2] He Zhang[2] Zhe Lin[2] Yuqian Zhou[2] Qing Liu[2] Shiwei Zhang[1] Yijun Li[2]
Shaoteng Liu[2] Haitian Zheng[2] Jason Kuen[2] Yuehuan Wang[1] Changxin Gao[1] Nong Sang[1]

[1]Key Laboratory of Image Processing and Intelligent Control,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Adobe Research

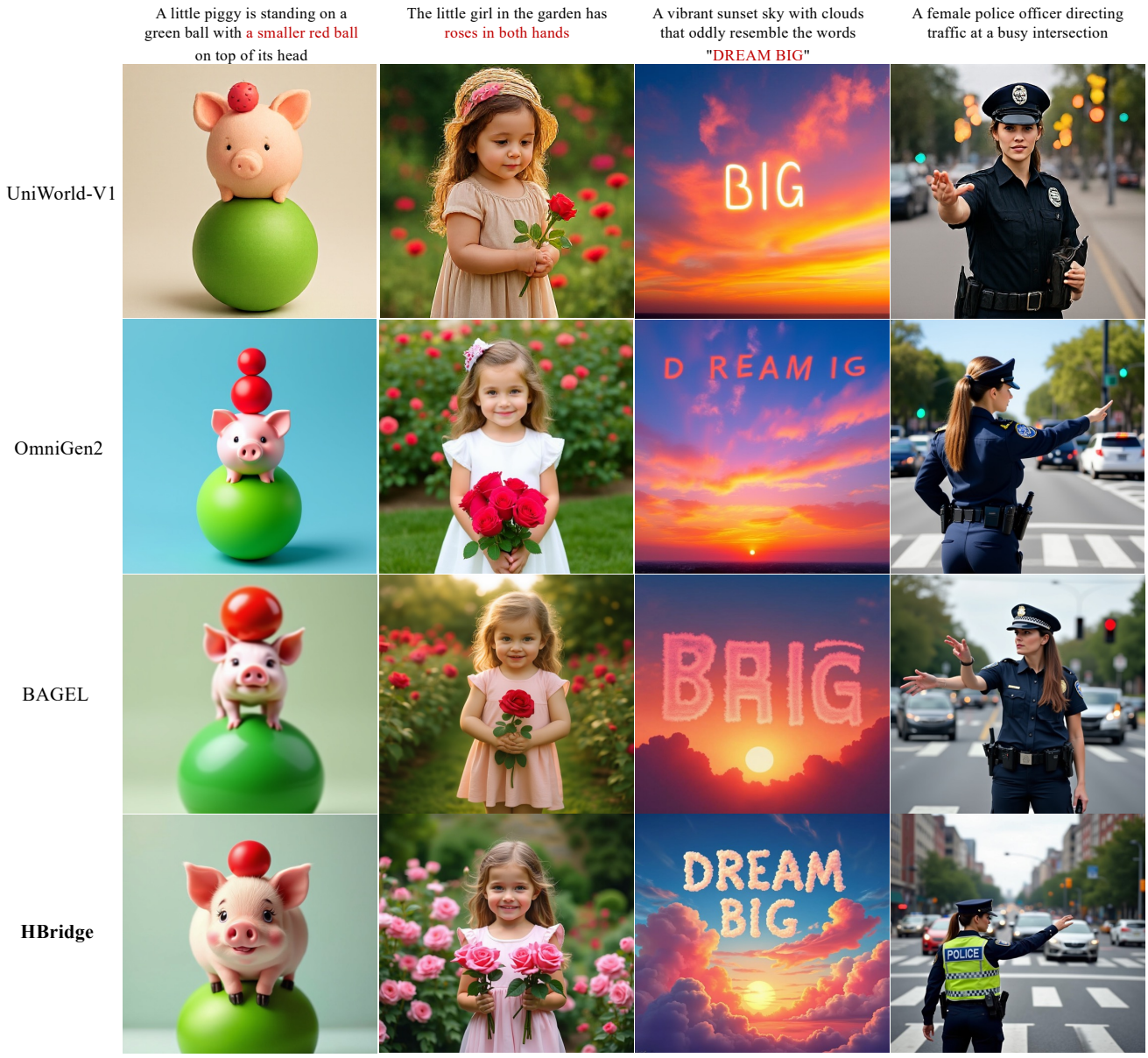

Figure 1. Comparison with the state-of-the-art methods such as UniWorld-V1, OmniGen2 and BAGEL on text-to-image generation task.

Figure 2. More qualitative results of text-to-image generation synthesized by the proposed `HBridge`.

Due to the page limit of the main document, we place some supplementary results and details in the appendix.

1. More Qualitative Results

In addition, we qualitatively compare our method with the state-of-the-art methods, including UniWorld-V1 [2], OmniGen2 [4], and BAGEL [1]. The text-to-image results are shown in Fig. 1. Our method demonstrates better semantic coherence and visual quality. As exhibited in Fig. 2, we show more high-quality, photorealistic text-to-image cases generated by `HBridge` with different resolution rates. These examples demonstrate excellent spatial layout, quantity control, and text rendering, validating powerful generative capabilities of `HBridge`. The editing results are displayed in Fig. 3, and `HBridge` can precisely understand and respond to the user's intentions, resulting in reliable editing outcomes. We attribute this to the fact that our H-shape design with semantic tokens helps improve the

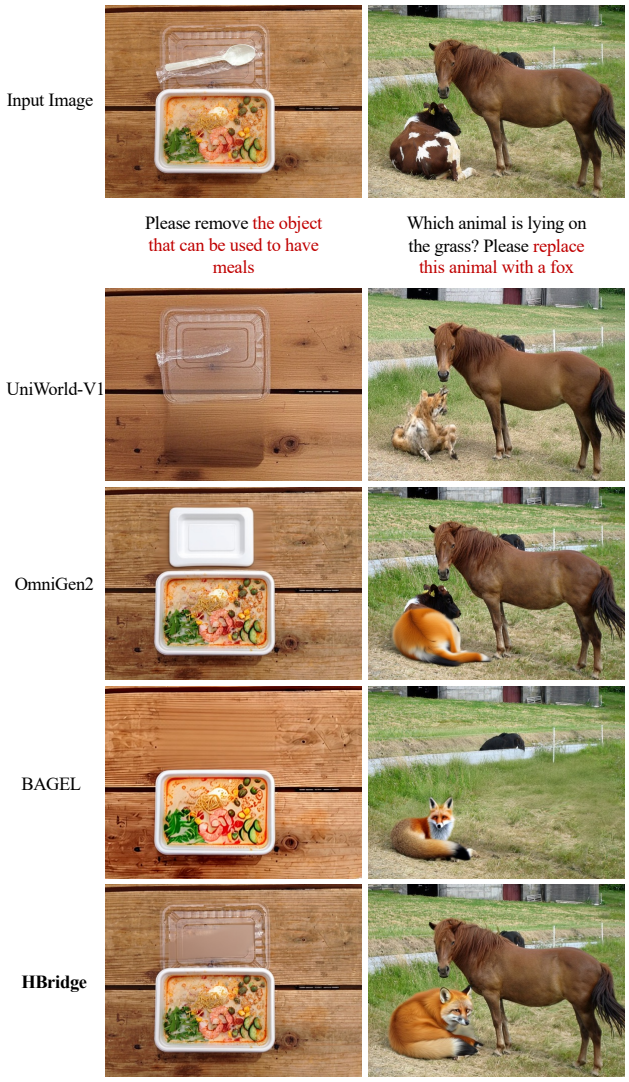

Figure 3. Comparison with state-of-the-art methods on image editing tasks, including object removal and replacement.

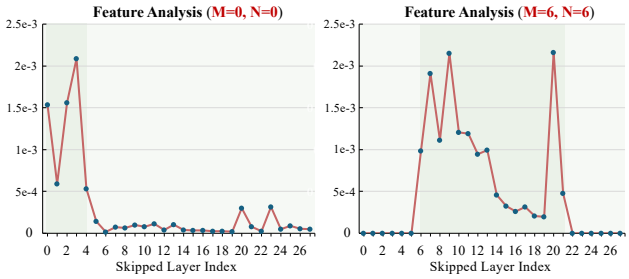

Figure 4. Analysis of varying the skipped layer under the **7B+4B** setting with 28 Transformer layers. We disconnect the multimodal self-attention layer by layer and analyze the differences in output features caused by disconnecting and reconnecting the multimodal self-attention layer. These differences are measured using the average normalized MSE [3] on DPG-Bench.

Table 1. Ablation study on the number of skipped layers under the **7B+4B** setting. LLM re-writer is not used on GenEval.

Setting	DPG-Bench	GenEval
M=0, N=0	83.21	0.80
M=6, N=6 **(Ours)**	**85.23**	**0.83**

Figure 5. Qualitative ablation study on the number of skipped layers under the **7B+4B** setting.

Table 2. Ablation study on the number of learnable semantic tokens under the 0.5B+4B setting.

Setting	DPG-Bench	GenEval
4 tokens	79.82	0.65
16 tokens	**80.03**	**0.66**
36 tokens	79.83	**0.66**

semantic understanding capabilities of generative models.

2. Additional Ablation Results

Overfitting Phenomenon under 7B+4B Settings. To further verify that the fully layer-by-layer connected method may easily overfit the shallow features of the understanding expert, we conduct additional experiments under the 7B+4B configuration. From the result in Fig. 4, it can be seen

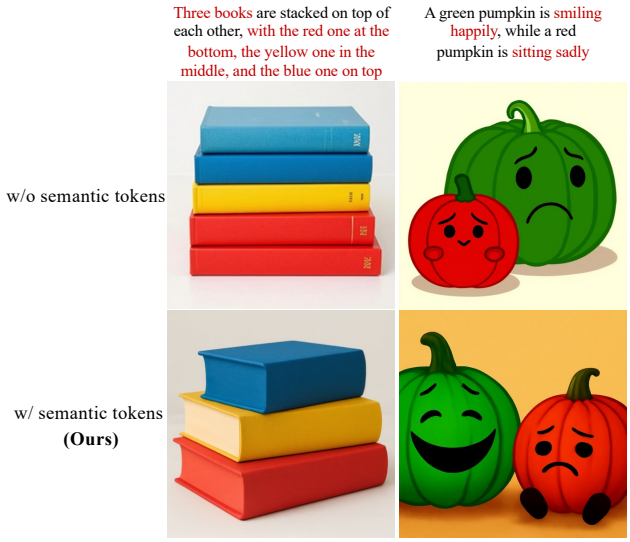

Figure 6. Qualitative ablation study on the effect of the proposed semantic reconstruction tokens.

that M=N=0 easily leads to overfitting of shallow features, while the setting of `HBridge` focuses primarily on the features of the intermediate semantic layers, resulting in better semantic coherence. We also show the quantitative results under the 7B+4B setting in Tab. 1, we can find that the mid-layer bridge (M=N=6) performs better than the baseline counterpart (M=N=0). In addition, the qualitative results in Fig. 5 demonstrate that M=N=0 may ignore some high-level semantics in textual prompts. The conclusions are consistent with those under the 0.5B+4B setting in the main document.

Effect of Semantic Reconstruction Tokens. To qualitatively analyze the efficacy of the proposed semantic reconstruction tokens, we visualize some examples in Fig. 6. From the results, we can notice that incorporating semantic reconstruction tokens helps to enhance the ability to perceive position and attributes. As shown in Tab. 2, we further conduct an ablation study on the number of learnable semantic tokens and find that 16 tokens achieve excellent performance.

References

[1] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2

[2] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 2

[3] Attilio A Poli and Mario C Cirillo. On the use of the normalized mean square error in evaluating dispersion model performance. *Atmospheric Environment. Part A. General Topics*, 27(15):2427–2434, 1993. 3

[4] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2