TALK2MOVE: Reinforcement Learning for Text-Instructed Object-Level Geometric Transformation in Scenes

Supplementary Material

Table 1. **Quantitative comparison** with Edit-R1 [3] on both **synthetic** and **real** benchmarks for object transformation tasks. Results reported in translation distance, rotation/scaling errors and editing accuracy.

Methods	Synthetic Benchmark						Real Benchmark					
	Translation		Rotation		Resize		Translation		Rotation		Resize	
	Trans. Dist.↑	Acc.↑	Rot. Err.↓	Acc.↑	Scale Err.↓	Acc.↑	Trans. Dist.↑	Acc.↑	Rot. Err.↓	Acc.↑	Scale Err.↓	Acc.↑
Edit-R1 [3]	**0.6762**	73.33%	0.2881	15.90%	0.3915	7.50%	0.5044	50.00%	0.2448	18.75%	0.8204	0.0%
Ours	0.6667	**76.67%**	**0.2861**	**29.55%**	**0.3894**	**9.17%**	**0.5196**	**53.85%**	**0.1997**	**31.25%**	**0.5947**	**7.14%**

A. Comparison with RL-based Image Editing

In this section, we compare TALK2MOVE with a concurrent RL-based image editing approach (released in October 2025), Edit-R1 [3], on object transformation tasks. Edit-R1 builds on the DiffusionNFT [7] pipeline and is trained on the eight standard editing tasks plus an additional red-box control task that is designed to move objects into a drawn red box. We evaluate on both synthetic and real image inputs. Our method achieves higher translation accuracy and produces larger average translation distances on real images. On synthetic inputs, Edit-R1 attains larger translation distances but suffers from reduced accuracy due to its weaker identity preservation.

We also provide qualitative comparisons in Fig. 1. When moving the knife into the white cup, Edit-R1 fails to preserve the objects originally inside the cup and does not maintain the surrounding background details well.

B. Unified Multi-task Training for SFT and GRPO

In this section, we analyze the effect of training the three transformation tasks—translation, rotation, and resizing—either separately or in a unified manner. From Tab. 2, we observe that single-task SFT yields worse performance than unified SFT, where samples from all three tasks are combined. This suggests that the three tasks benefit from mutual learning during supervised fine-tuning.

In the GRPO training, however, starting from the unified SFT checkpoint, we find that training each task individually performs slightly better than continuing GRPO on the combined multi-task setting. Simply merging the three tasks for unified GRPO causes the learned rollout distribution to be biased toward performing better on translation, while providing limited gains on rotation and resizing. A possible explanation is that GRPO primarily focuses on unlocking the model's hidden capabilities by learning a robust answer ranking. When applied to a multi-task GRPO setting, the complexity of learning a consistent ranking across heterogeneous tasks increases substantially, making single-task GRPO more effective.

Table 2. **Ablative Study** on unified SFT and GRPO training for object transformation tasks.

Methods	Translation		Rotation		Resize	
	Trans. Dist. ↑	Acc. ↑	Rot. Err. ↓	Acc. ↑	Scale Err. ↓	Acc. ↑
Single Task SFT	0.5948	65.71%	0.2979	13.64%	0.3899	7.50%
Single Task GRPO	**0.6667**	73.13%	0.2969	27.91%	0.3912	7.50%
Unified SFT	0.6416	75.00%	0.3197	25.00%	0.3917	7.50%
Unified GRPO	0.6486	**76.67%**	0.2880	25.58%	**0.3805**	5.83%
Ours	**0.6667**	**76.67%**	**0.2861**	**29.55%**	0.3894	**9.17%**

C. Robustness and Consistency of Reward Variance

We investigate the stability of reward variance across two dimensions: cross-sample consistency and temporal stability during policy updates.

Cross Sample Consistency. We analyze whether the optimal exit step is tied to specific image structures or remains consistent across a task domain. As shown in Fig. 2, for tasks like image translation, the reward variance consistently peaks at step 4 regardless of the individual image content. This per-task consistency suggests that learning bottlenecks are inherent to the task difficulty at certain diffusion stages rather than sample-specific noise. **Temporal Stability during Policy Updates.** We also verify if the reward variance distribution estimated via off-policy evaluation remains a valid proxy as the policy model evolves. This question is

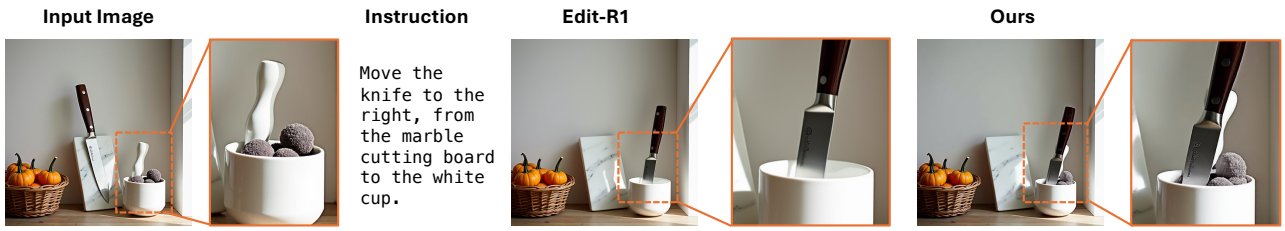

Figure 1. **Additional Qualitative Comparison** with Edit-R1 [3] on the object translation task. Our TALK2MOVEis better at keeping the original scene details while performing effective object translation.

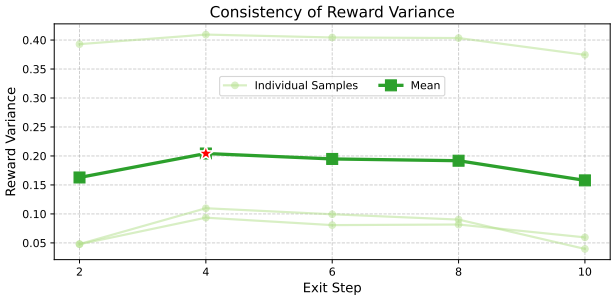

Figure 2. **Cross Sample Consistency** of Reward Variance Distribution on object translation task.

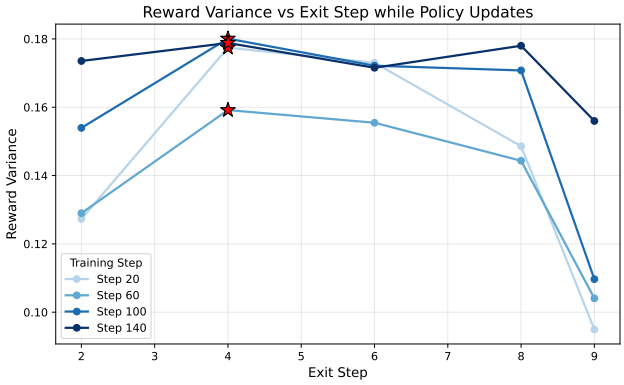

Figure 3. **Temporal Stability** of Reward Variance Distribution during full sampling policy updates on object translation task. The red star indicates the maximum reward variance.

important because we rely on reward variance as an indirect measure of sampling efficiency for selecting informative steps, and this pre-computed metric must remain stable throughout training to be valid.

To assess this, we first train the policy in a setting where all steps are perturbed, and then recompute the reward-variance distribution using multiple intermediate checkpoints, as illustrated in Fig. 3. For each checkpoint, the variance is estimated using 128 rollouts per sample and averaged over two randomly chosen samples. The x-axis denotes the exit step, and the y-axis shows the corresponding reward variance. Curve colors range from light to dark, representing checkpoints from early to late stages of training.

The results show that the distributions remain largely stable over the course of training. In particular, exiting at step 4 consistently exhibits the highest variance, indicating that perturbations at this step yield the greatest sampling efficiency. This consistency demonstrates that off-policy evaluation provides a reliable signal for identifying informative steps in our active-sampling scheme.

D. Analysis on Shortcut Connections in GRPO

In this section, we investigate the effect of applying shortcuts at different diffusion steps. In Fig. 4, we observe that applying a one-step shortcut at early diffusion steps leads to suboptimal denoising quality, which in turn reduces the robustness of reward estimation. However, using a two-step ODE shortcut effectively mitigates this issue and provides stable denoising performance even in early stages. In later diffusion steps, a one-step ODE shortcut already produces results comparable to full-sample denoising.

We also find that perturbing too many steps introduces additional noise, especially in the late denoising stages where the model is primarily responsible for removing fine-grained noise. Injecting perturbations at this point can degrade image quality rather than improve it. These findings indicate that shortcut-based acceleration is an effective strategy for alleviating the long sampling time required during GRPO online rollouts.

E. Extended Discussion on Prompt Diversity

A potential concern in our text-guided geometric manipulation system is whether the performance gains are tied to specific prompt templates used during training, which might disadvantage baseline models. To assess this *prompt sensitivity*, we conducted a robust evaluation by comparing the original task prompts against four diverse recaptioned variants that explicitly specify object-centric rotations with varying linguistic structures.

Experimental results demonstrate that both our proposed method ($\sigma_{Acc} = 0.0324, \sigma_{err} = 0.0113$) and the QwenImageEdit baseline ($\sigma_{Acc} = 0.0410, \sigma_{err} = 0.0136$) exhibit

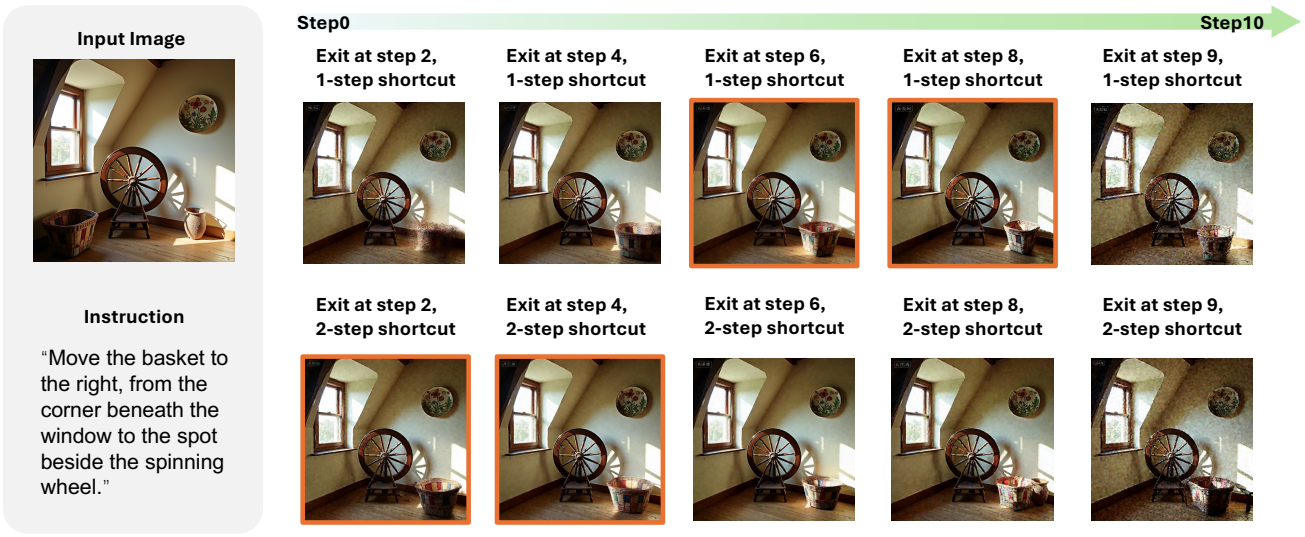

Figure 4. **Denoising results with Shortcut connections** in active step sampling.

minimal performance variance across different prompt formulations. The consistent metric stability suggests that the observed improvements stem from the model's enhanced geometric transformation ability, rather than an over-reliance on template-specific cues.

F. Extended Discussion on Consistency Metrics

While our current evaluation framework employs CLIP score and L_1 distance as primary baselines, we recognize that image consistency is a multi-dimensional concept that necessitates a more holistic and adaptive assessment.

We acknowledge that pixel-level metrics like L_1 distance can be overly stringent due to the inherent stochasticity of diffusion-based generative models. Such metrics may penalize valid structural variations that do not compromise overall visual quality. Similarly, as highlighted in DreamBench++ [5], CLIP-based evaluation, while effective for global semantic grounding, often lacks the granularity required to preserve fine-grained object identity during complex transformations. In this study, these metrics serve as a foundational baseline to demonstrate the fundamental efficacy and stability of our approach under strict constraints.

In future work, we will continue to explore VLM-based reward models that internalize human-aligned consistency priors, providing high-fidelity feedback for RL-based optimization to achieve more sophisticated and stable scene-level manipulations.

G. More Qualitative Results

We present additional qualitative examples comparing TALK2MOVE with several baseline methods, including Bagel [1], Flux-Kontext [2], GPT-Image-1 [4], QwenImageEdit [6], and Edit-R1 [3]. The models are tested on both synthetic and real image inputs, where we use blue color to indicate synthetic images and orange color to indicate real images. As shown in Figs. 5 to 7, baseline methods commonly exhibit issues such as poor preservation of scene details or object identity, duplicated objects when performing translation, unintended viewpoint rotations when attempting to rotate objects, and inappropriate scaling factors—either shrinking the object too much or enlarging it excessively—during resizing. In contrast, TALK2MOVE produces more faithful and controlled edits across all transformation types.

References

[1] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 3

[2] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 3

[3] Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, and Li Yuan. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025. 1, 2, 3

[4] OpenAI. Gpt-image-1: Openai's image generation model.

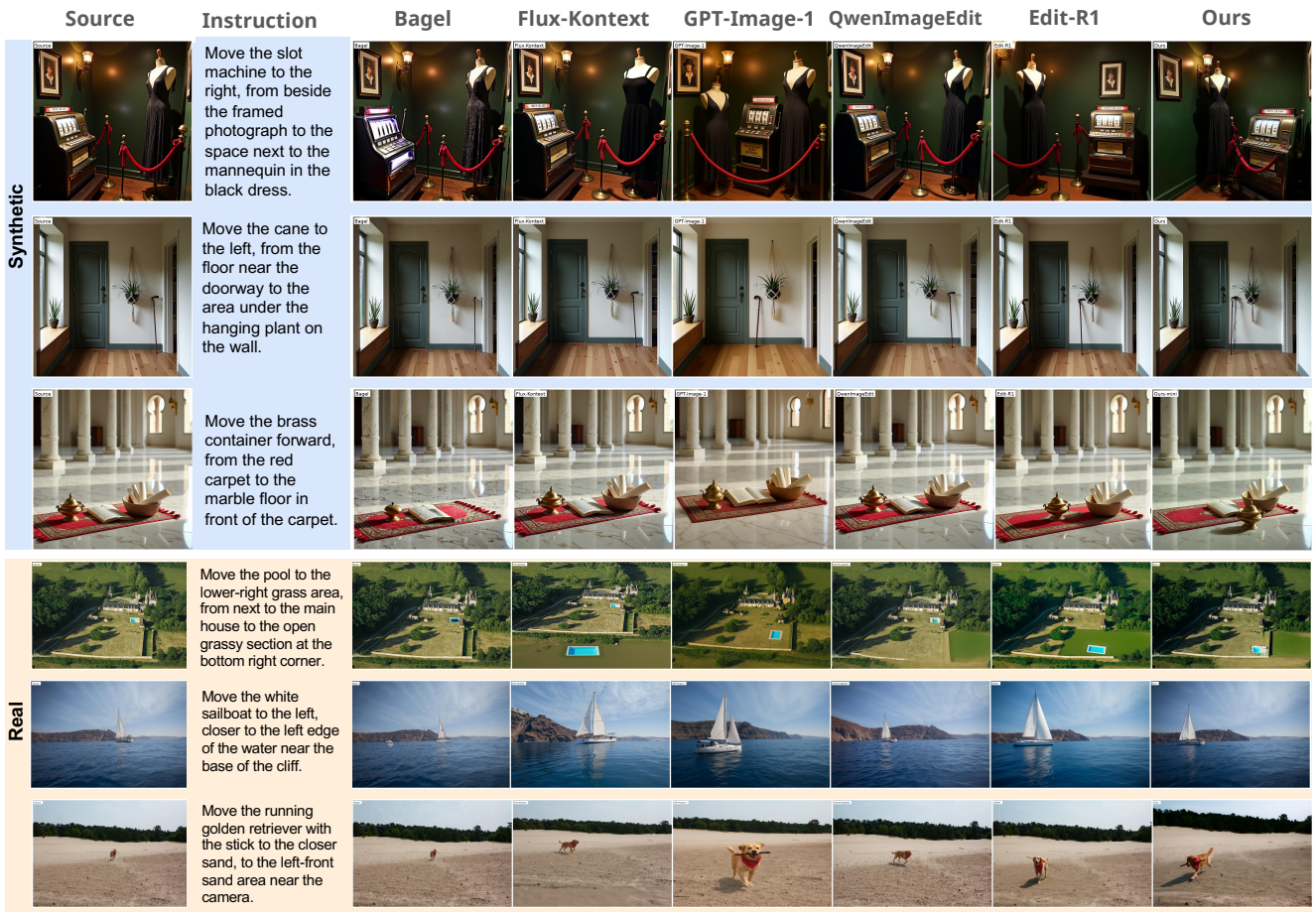

Figure 5. **Additional qualitative comparison** for the object translation task. We present qualitative results on both synthetic (blue) and real (orange) images to compare the object translation performance.

https://platform.openai.com/docs/models/gpt-image-1, 2024. Accessed: 2025-11-11. 3

[5] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 3

[6] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 3

[7] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025. 1

Figure 6. **Additional qualitative comparison** for the object rotation task. We present qualitative results on both synthetic (blue) and real (orange) images to compare the object rotation performance.

Figure 7. **Additional qualitative comparison** for the object resizing task. We present qualitative results on both synthetic (blue) and real (orange) images to compare the object resizing performance.