Supplementary Material of *Toward Better Out-painting: Improving the Image Composition with Initialization Policy Model*

Xuan Han[1] Yihao Zhao[1] Yanhao Ge[1,2] Mingyu You[1*]
[1] Tongji University, Shanghai, China [2] Future Imaging Area
{hanxuan, zhaoyihao, myyou}@tongji.edu.cn {halege}@vivo.com

1. Dataset Construction

In this section, we will provide a detailed description of the construction of the dataset. The entire pipeline consists of four steps:

- **Multi-modal Annotation** - Using BLIP2 [2] generates natural language captions for images. Using RAM [5] extracts the visual tags for images.
- **Semantic Filtering** - Llama3 [4] is used to filter the tags of focused object according to the caption and tag collection. The tags with obscure or background semantic will be removed, such as "close-up", "sky", "road", etc.
- **Box Generation** - GroundingDINO[3] predicts bounding boxes for filtered tags using caption-guided detection.
- **Mask Refinement** - HQ-SAM[1] produces high-quality masks within detected boxes, Masks that occupy too small a proportion in the image or have a low confidence level will be removed, with the thresholds being 0.1 and 0.35 respectively.

The test set used in this work is consists of 250 cases, among which 100 are validation images from OpenImage and 150 are high-quality images from web sources. Each image is equipped with manual foreground mask annotations and the correspond text prompt.

2. Hyperparam. Study of Reward Supervision

$\mathcal{L}_{rwd}$ plays an important role in the method proposed in this paper. The application of it involves two important hyperparameters, the sampling step K and the loss weight λ_{rwd}. Since the main focus of this method is on the image quality in the composition domain, we determine their values based on the SAM and SA metrics. The candidate values of K are $[5, 10, 20]$, while the candidate values of λ_{rwd} are $[0.1, 1.0, 10.0]$. The quantitative evaluation results are shown in Tab. 1.

It can be seen that when $K \in [10, 20]$, the performance of the SAM and SA metrics is relatively stable. Considering that increasing K will significantly increase the training time, we choose the smaller value among them. For the selection of λ_{rwd}, when it is set to a relatively small value, the effect of the reward supervision is not strong, and the model will additionally generate artifacts. When it is set too large, the problem of artifacts is alleviated, but the image quality significantly deteriorates. As introduced in Section 3.4, since the gradient of $\mathcal{L}_{rwd}$ is simplified, there will be certain errors in it. When λ_{rwd} is too large, the gradient with excessive fluctuations will have a negative impact on the training of IPM. Finally, we select $K = 10$ and $\lambda_{rwd} = 1$.

Value	K			λ_{rwd}		
	5	10	20	0.1	1.0	10.0
SAM $\downarrow$	0.117	**0.070**	<u>0.072</u>	0.089	<u>0.070</u>	**0.057**
AS $\uparrow$	5.29	**5.34**	<u>5.34</u>	<u>5.29</u>	**5.34**	4.96

Table 1. The quantitative evaluation results of hyperparameter study of reward supervision.

3. Hybrid Sampling Strategy

In Section 4.4, we noted that IPM sacrifices certain diversity to achieve more rational results. However, practical applications may require greater output diversity. Hybrid sampling serves as an alternative strategy in such scenarios. It samples x_{t^s} through a weighted combination of IPM's prediction and the original denoising process result $x_{t^s}^{ori} \sim p_\theta(x_{t^s}^{ori}|x_T, y, c)$, formulated as: $x_{t^s}^{hybrid} = a\,\phi(y, c) + (1 - a)\,x_{t^s}^{ori}, \quad 0 \leq a \leq 1$. In the following section, we will present evaluation results of this strategy and illustrate the trade-off between compositional rationality and image diversity as weight a is adjusted.

Fig. 3 visually demonstrates the potential improvements and limitations of this strategy. In the hot air balloon example (good case), smaller a values introduce more color tones, addressing the grayish hue issue when using pure IPM ($a = 1.0$). Conversely, in the pumpkin example (bad case), reducing a leads to severe artifacts and unrealistic vi-

*Corresponding Author. M. You is with the College of Electronic and Information Engineering, Tongji University, State Key Laboratory of Autonomous Intelligent Unmanned Systems

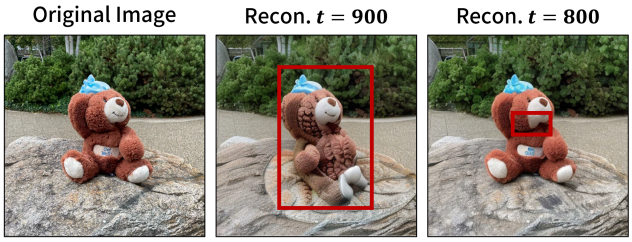

Figure 1. The deviations introduced by the inversion operation.

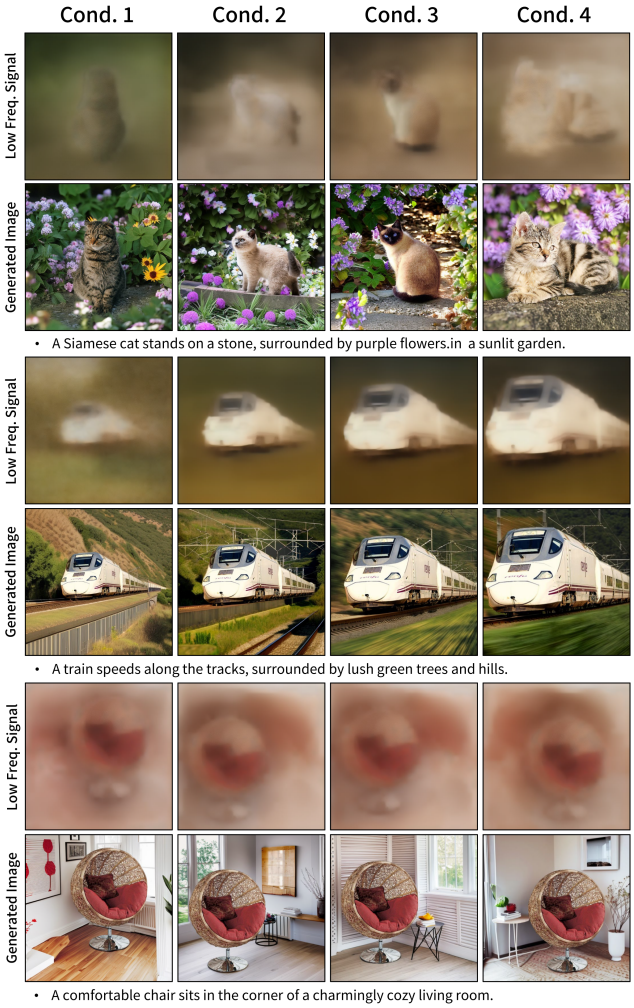

Figure 2. The generated images and low-frequency signals (predicted by IPM) of foreground objects with variable identities, scales, and positions.

sual effects. In summary, hybrid sampling strategy offers a way to compensate for this limitation and achieve a balance between compositional rationality and image diversity.

4. More Visual Results

In Section 3.4, we pointed out that the inversion operation may introduce deviations in the reconstruction results. Fig. 1 provides an intuitive illustration. The red boxes highlight the regions where significant deviations occur. It can be observed that the earlier the time step of inversion, the more intense the deviation. This result once again confirms the necessity of using online diffusion reward supervision.

In Section 4.3, we presented the generated images of foreground objects with variable identities, scales, and positions. Fig. 2 shows the corresponding low-frequency signals predicted by IPM.

Due to space limitations, only a limited number of results are presented in the main text to demonstrate the adaptability of IPM to different foreground objects and text prompts. This part provides more visual illustrations. Fig. 4 shows additional results regarding the foreground objects with varying scales and positions. Fig. 5 presents more results concerning the diverse text prompts. The results without using IPM are also included as a comparison to better showcase the advantages of the method proposed in this paper.

References

[1] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, pages 29914–29934, 2024. 1

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 1

[3] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499, 2023. 1

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and Faisal Azhar. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 1

[5] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize anything: A strong image tagging model. In *CVPR*, pages 1724–1732, 2024. 1

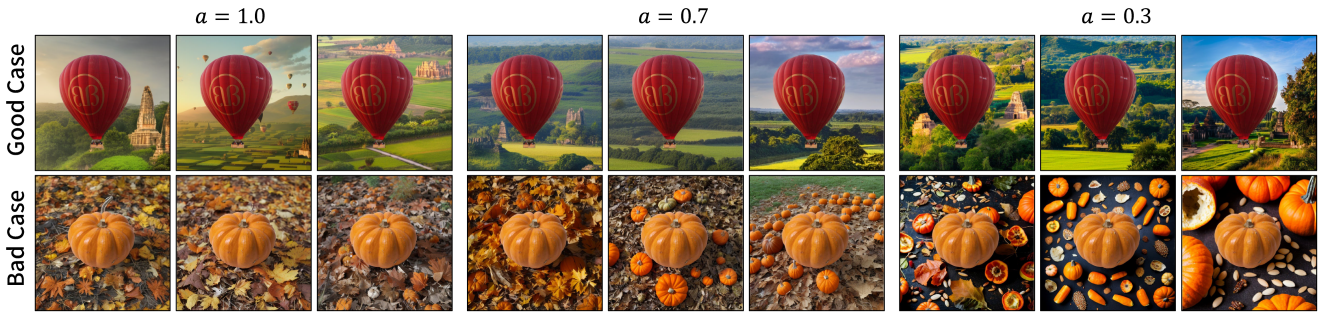

Figure 3. The good case and bad case of hybrid sampling strategy.

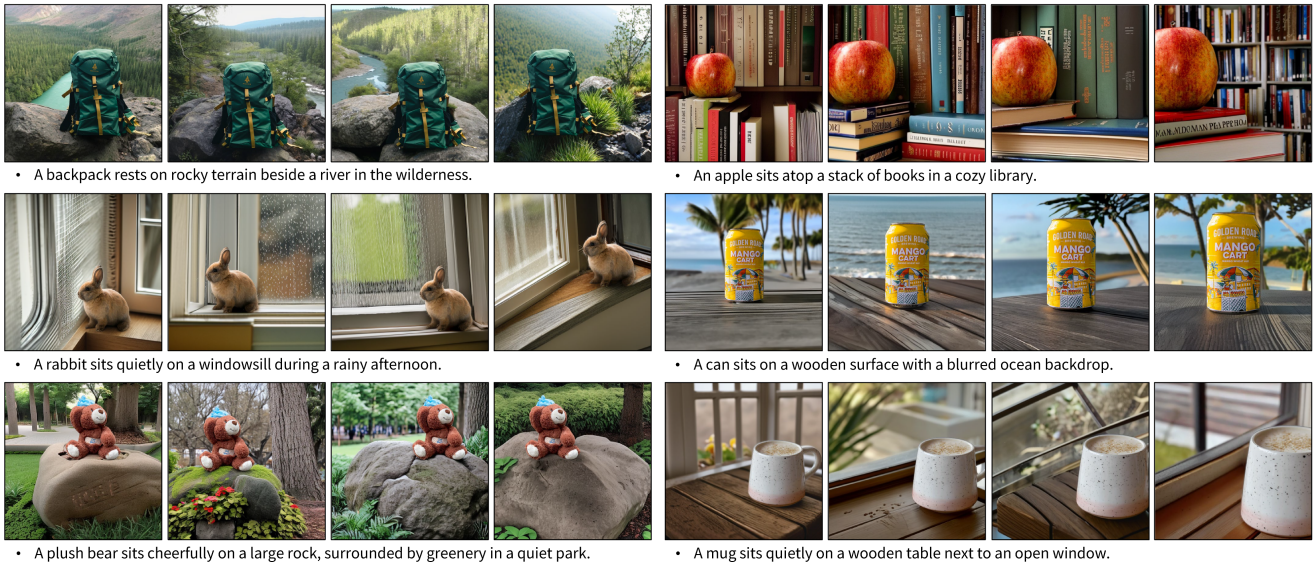

Figure 4. More samples of generated images of foreground object with variable identities, scales, and positions.

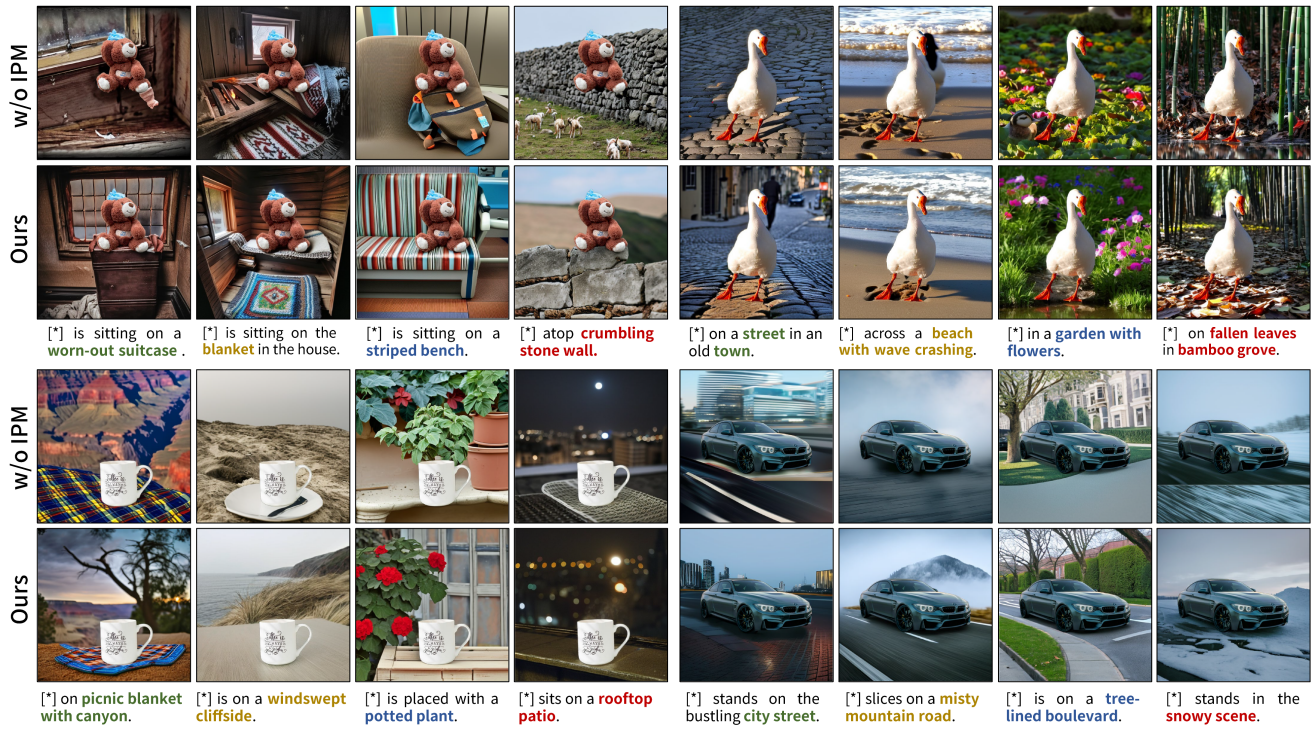

Figure 5. More samples of generated images of the baseline model with or w/o IPM under different text prompts.