FreeFlux: Understanding and Exploiting Layer-Specific Roles in RoPE-Based MMDiT for Versatile Image Editing-Supplementary Material

Tianyi Wei[1], Yifan Zhou[1], Dongdong Chen[2], Xingang Pan[1]
[1]S-Lab, Nanyang Technological University [2]Microsoft GenAI
{tianyi.wei, yifan006, xingang.pan}@ntu.edu.sg , cddlyf@gmail.com
https://wtybest.github.io/projects/FreeFlux/

1. Algorithm

The pseudo-code for performing object addition, non-rigid editing, background replacement, object movement, and outpainting using our method is provided in Algorithms 1, 2, 3, 4, and 5.

2. Generalization to Lumina-Next

We validate the generalization of our method on Lumina-Next [5], a RoPE-based DiT model with alternating self- and cross-attention layers. By applying our probing technique, we identify the most position-sensitive layers [10,46,14,6,28,20] and the most content-similarity-sensitive layers [0,2,18,42,38,4], as visualized in Figure 1. Using these selected layers, we perform object addition ("Sail") and non-rigid editing ("Running") tasks, respectively, and compare them against the baseline that uses all layers (see Figure 2). These results strongly support the generalization ability of our approach to new architectures.

Figure 1. Layer-wise positional dependency in Lumina-Next.

Figure 2. Customized editing results with Lumina-Next.

3. More Qualitative Results

In Figures 3, 4, 5, and 6 we give more visual comparison results with other methods and our results on the outpainting and object moving tasks.

References

[1] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3, 4

[2] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024. 2, 3, 4

[3] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 2, 3, 4

[4] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023. 2, 3, 4

[5] Le Zhuo, Ruoyi Du, Xiao Han, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024. 1

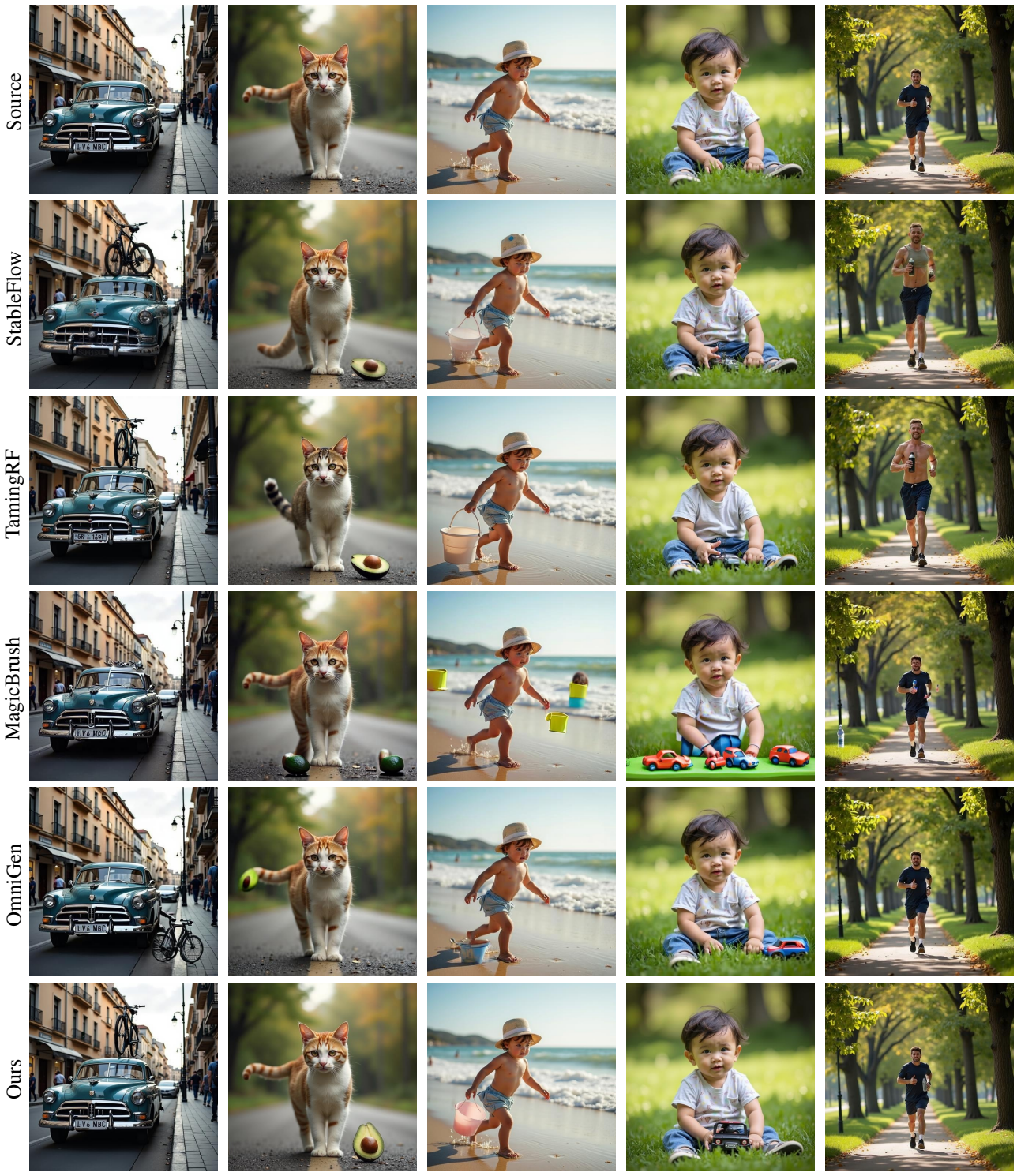

Figure 3. Qualitative comparison on the object addition task with training-free methods StableFlow [1] and TamingRF [2], as well as general image editing models MagicBrush [4] and OmniGen [3]. Our method not only achieves high-quality editing results but also demonstrates the best ability to preserve irrelevant regions.

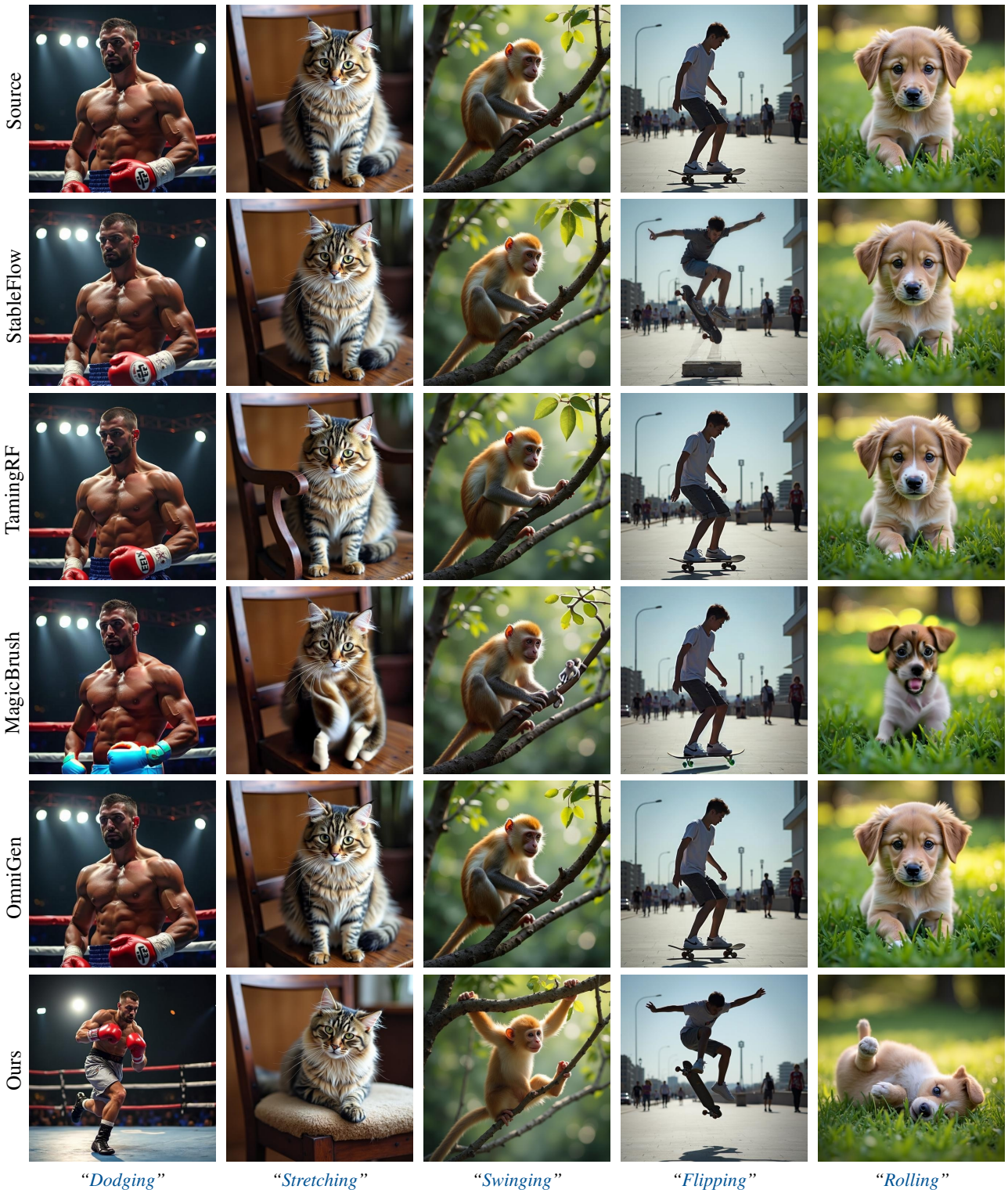

Figure 4. Qualitative comparison on the non-rigid editing task with training-free methods StableFlow [1] and TamingRF [2], as well as general image editing models MagicBrush [4] and OmniGen [3]. Our method effectively balances object deformation and appearance transfer.

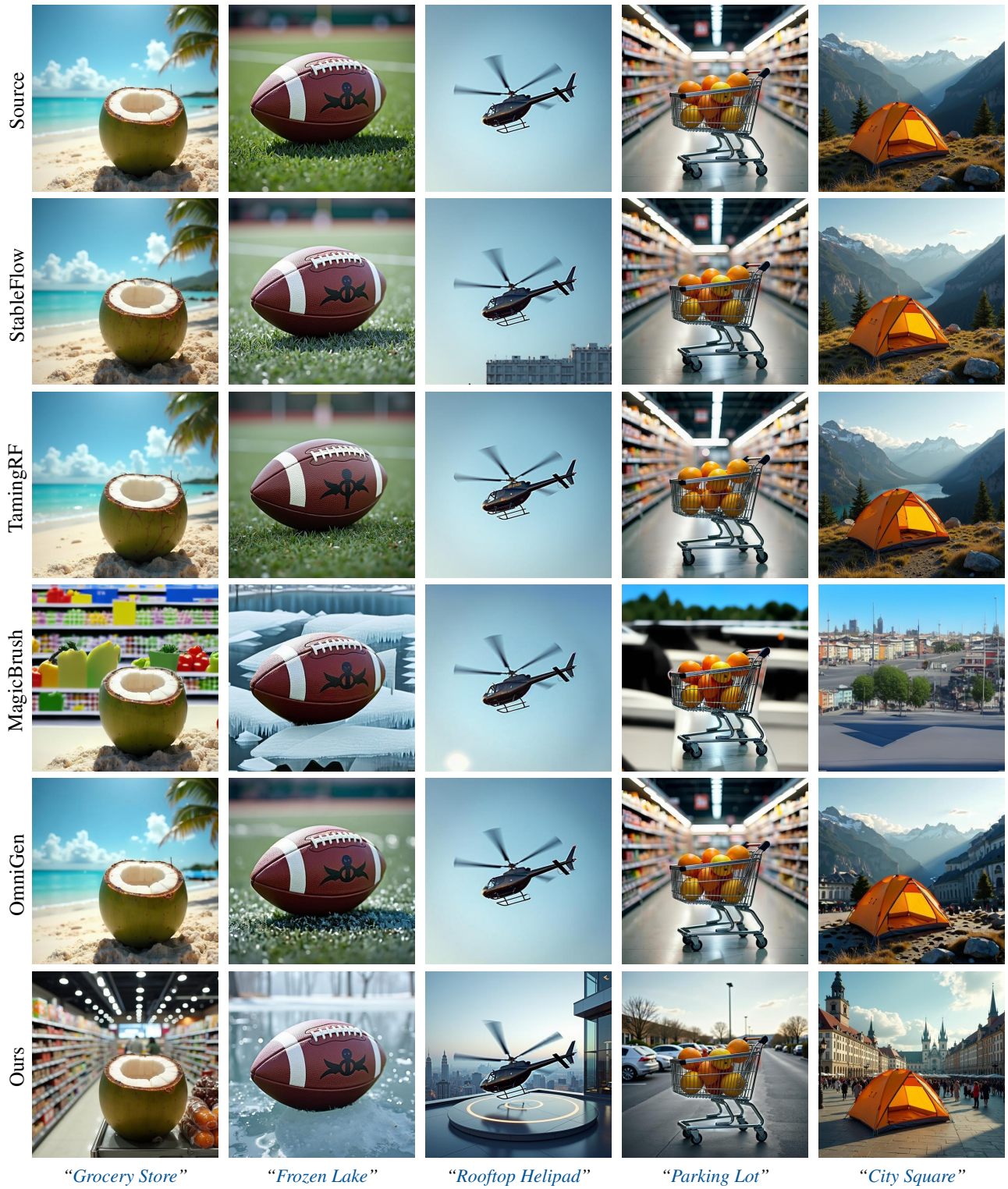

Figure 5. Qualitative comparison on the background replacement task with training-free methods StableFlow [1] and TamingRF [2], as well as general image editing models MagicBrush [4] and OmniGen [3]. Our approach delivers the most visually compelling background changes while preserving the foreground object intact.

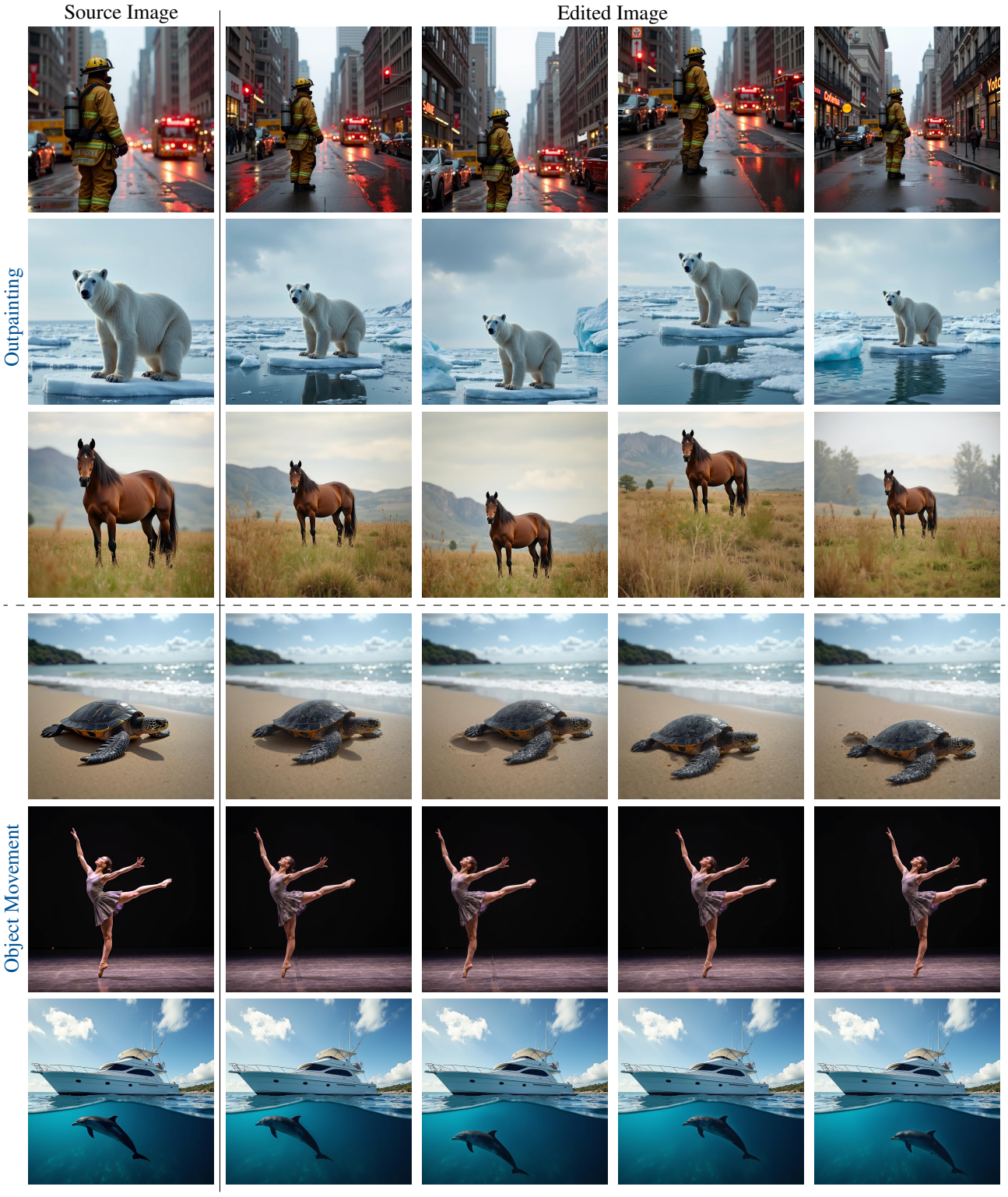

Figure 6. Visual results of our method on region-preserved editing tasks such as object movement and outpainting.

Algorithm 1: Object Addition

Input: A source prompt $\mathcal{P}_{src}$, A editing prompt $\mathcal{P}_{edit}$, a pretrained RoPE-based MMDiT text-to-image model ε_θ, an image decoder $\mathcal{D}$, total sampling steps T (default 50), denoising step R (default 7) when applying Reasoning-before-Generation, the most position-dependent layers $\mathbb{P}$.

Output: An edited image x^{edit} aligned with the editing prompt $\mathcal{P}_{edit}$.

Initialize $z_T^{src} \sim \mathcal{N}(0,1)$, $z_T^{edit} \leftarrow z_T^{src}$, $z_{init} \leftarrow z_T^{src}$.
for $t = T$ *to* $T - R + 1$ **do**
  for i *in* $\mathbb{P}$ **do**
    $Q_{src}^{t-i}, K_{src}^{t-i}, V_{src}^{t-i} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
    $Q_{edit}^{t-i}, K_{edit}^{t-i}, V_{edit}^{t-i} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit})$
    $Attn(Q_{edit}^{t-i}, K_{src}^{t-i}, V_{src}^{t-i})$
  end
  $z_{t-1}^{src} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
  $z_{t-1}^{edit} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit}, Attn)$
end
The added object region mask M_{obj} is reasoned out.
$z_T^{src} \leftarrow z_{init}$, $z_T^{edit} \leftarrow z_{init}$
for $t = T$ *to* 1 **do**
  for i *in* $\mathbb{P}$ **do**
    $Q_{src}^{t-i}, K_{src}^{t-i}, V_{src}^{t-i} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
    $Q_{edit}^{t-i}, K_{edit}^{t-i}, V_{edit}^{t-i} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit})$
    $K_{obj}^{t-i} = M_{obj} \times K_{edit}^{t-i} + (1 - M_{obj}) \times K_{src}^{t-i}$
    $V_{obj}^{t-i} = M_{obj} \times V_{edit}^{t-i} + (1 - M_{obj}) \times V_{src}^{t-i}$
    $Attn(Q_{edit}^{t-i}, K_{obj}^{t-i}, V_{obj}^{t-i})$
  end
  $z_{t-1}^{src} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
  $z_{t-1}^{edit} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit}, Attn)$
end
$x^{src} \leftarrow \mathcal{D}(z_0^{src})$
$x^{edit} \leftarrow \mathcal{D}(z_0^{edit})$
Return: x^{edit}

Algorithm 2: Non-Rigid Editing

Input: A source prompt $\mathcal{P}_{src}$, A editing prompt $\mathcal{P}_{edit}$, a pretrained RoPE-based MMDiT text-to-image model ε_θ, an image decoder $\mathcal{D}$, total sampling steps T (default 50), the more content-similarity-dependent layers $\mathbb{C}$.

Output: An edited image x^{edit} aligned with the editing prompt $\mathcal{P}_{edit}$.

Initialize $z_T^{src} \sim \mathcal{N}(0,1)$, $z_T^{edit} \leftarrow z_T^{src}$.
for $t = T$ *to* 1 **do**
  for i *in* $\mathbb{C}$ **do**
    $Q_{src}^{t-i}, K_{src}^{t-i}, V_{src}^{t-i} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
    $Q_{edit}^{t-i}, K_{edit}^{t-i}, V_{edit}^{t-i} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit})$
    $Attn(Q_{edit}^{t-i}, K_{src}^{t-i}, V_{src}^{t-i})$
  end
  $z_{t-1}^{src} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
  $z_{t-1}^{edit} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit}, Attn)$
end
$x^{src} \leftarrow \mathcal{D}(z_0^{src})$
$x^{edit} \leftarrow \mathcal{D}(z_0^{edit})$
Return: x^{edit}

Algorithm 3: Background Replacement

Input: A source prompt $\mathcal{P}_{src}$, A editing prompt $\mathcal{P}_{edit}$, a pretrained RoPE-based MMDiT text-to-image model ε_θ, an image decoder $\mathcal{D}$, total sampling steps T (default 50), denoising step B (default 45) to stop value blending, total number of layers L, the foreground mask automatically derived by SAM2 M_{fg}^{sam}.

Output: An edited image x^{edit} aligned with the editing prompt $\mathcal{P}_{edit}$.

Initialize $z_T^{src} \sim \mathcal{N}(0,1)$, $z_T^{edit} \leftarrow z_T^{src}$.
for $t = T$ *to* $T - B + 1$ **do**
  for $i = 1$ *to* L **do**
    $Q_{src}^{t-i}, K_{src}^{t-i}, V_{src}^{t-i} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
    $Q_{edit}^{t-i}, K_{edit}^{t-i}, V_{edit}^{t-i} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit})$
    $V_{fg}^{t-i} = M_{fg}^{sam} \times V_{src}^{t-i} + (1 - M_{fg}^{sam}) \times V_{edit}^{t-i}$
    $Attn(Q_{edit}^{t-i}, K_{edit}^{t-i}, V_{fg}^{t-i})$
  end
  $z_{t-1}^{src} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
  $z_{t-1}^{edit} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit}, Attn)$
end
for $t = T - B$ *to* 1 **do**
  $z_{t-1}^{src} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
  $z_{t-1}^{edit} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit})$
end
$x^{src} \leftarrow \mathcal{D}(z_0^{src})$
$x^{edit} \leftarrow \mathcal{D}(z_0^{edit})$
Return: x^{edit}

Algorithm 4: Object Movement

Input: A source prompt $\mathcal{P}_{src}$, A editing prompt $\mathcal{P}_{edit}$, a pretrained RoPE-based MMDiT text-to-image model ε_θ, an image decoder $\mathcal{D}$, total sampling steps T (default 50), denoising step B (default 45) to stop value blending, total number of layers L, the coordinate c of the movement direction, the function MAP that maps the source object value to a specified location based on c and copies the unaffected region.

Output: An edited image x^{edit} aligned with the editing prompt $\mathcal{P}_{edit}$.

Initialize $z_T^{src} \sim \mathcal{N}(0,1)$, $z_T^{edit} \leftarrow z_T^{src}$.
for $t = T$ *to* $T - B + 1$ **do**
  for $i = 1$ *to* L **do**
    $Q_{src}^{t-i}, K_{src}^{t-i}, V_{src}^{t-i} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
    $Q_{edit}^{t-i}, K_{edit}^{t-i}, V_{edit}^{t-i} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit})$
    $V_{move}^{t-i} = MAP(V_{edit}^{t-i}, V_{src}^{t-i}, c)$
    $Attn(Q_{edit}^{t-i}, K_{edit}^{t-i}, V_{move}^{t-i})$
  end
  $z_{t-1}^{src} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
  $z_{t-1}^{edit} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit}, Attn)$
end
for $t = T - B$ *to* 1 **do**
  $z_{t-1}^{src} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
  $z_{t-1}^{edit} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit})$
end
$x^{src} \leftarrow \mathcal{D}(z_0^{src})$
$x^{edit} \leftarrow \mathcal{D}(z_0^{edit})$
Return: x^{edit}

Algorithm 5: Outpainting

Input: A source prompt $\mathcal{P}_{src}$, A editing prompt $\mathcal{P}_{edit}$, a pretrained RoPE-based MMDiT text-to-image model ε_θ, an image decoder $\mathcal{D}$, total sampling steps T (default 50), denoising step B (default 45) to stop value blending, total number of layers L, the paste coordinates c of the original image on the higher-resolution edited image, the function $PASTE$ copies the value of the original image to the corresponding position in the edited image based on c.

Output: An edited image x^{edit} aligned with the editing prompt $\mathcal{P}_{edit}$.

Initialize $z_T^{src} \sim \mathcal{N}(0,1)$, $z_T^{edit} \leftarrow z_T^{src}$.
for $t = T$ *to* $T - B + 1$ **do**
  for $i = 1$ *to* L **do**
    $Q_{src}^{t-i}, K_{src}^{t-i}, V_{src}^{t-i} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
    $Q_{edit}^{t-i}, K_{edit}^{t-i}, V_{edit}^{t-i} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit})$
    $V_{out}^{t-i} = PASTE(V_{edit}^{t-i}, V_{src}^{t-i}, c)$
    $Attn(Q_{edit}^{t-i}, K_{edit}^{t-i}, V_{out}^{t-i})$
  end
  $z_{t-1}^{src} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
  $z_{t-1}^{edit} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit}, Attn)$
end
for $t = T - B$ *to* 1 **do**
  $z_{t-1}^{src} \leftarrow \varepsilon_\theta(z_t^{src}, t, \mathcal{P}_{src})$
  $z_{t-1}^{edit} \leftarrow \varepsilon_\theta(z_t^{edit}, t, \mathcal{P}_{edit})$
end
$x^{src} \leftarrow \mathcal{D}(z_0^{src})$
$x^{edit} \leftarrow \mathcal{D}(z_0^{edit})$
Return: x^{edit}