APPENDIX

A PROOF OF PROPOSITION 4.1

Prop. 4.1: *Suppose the velocity field $\boldsymbol{v}_\theta$ is Lipschitz, and there is a constant C such that $\left\|\boldsymbol{Z}_{t_p}-\boldsymbol{Z}_{t_q}\right\| \leq C\left\|t_p-t_q\right\|, \forall t_p, t_q \in [0,1]$, where $\boldsymbol{Z}_{t_p}$ and $\boldsymbol{Z}_{t_q}$ come from the same sampling process. Then for any two consecutive steps t_{i-1} and t_i, the local error of inversion and reconstruction using Uni-Inv is $\mathcal{O}\left(\Delta t_i^3\right)$, where $\Delta t_i = t_i - t_{i-1}$.*

Assumption 1. *The velocity function $\boldsymbol{v}_\theta(\cdot,\tau)$ is X_1-Lipschitz for $\forall\ \tau \in [0,1]$,* i.e., *given a τ, $\|\boldsymbol{v}_\theta(\boldsymbol{\zeta}_1,\tau)-\boldsymbol{v}_\theta(\boldsymbol{\zeta}_2,\tau)\| \leq X_1\|\boldsymbol{\zeta}_1-\boldsymbol{\zeta}_2\|$ for $\forall\ \boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2$.*

Assumption 2. *The velocity function $\boldsymbol{v}_\theta(\boldsymbol{\zeta},\cdot)$ is X_2-Lipschitz for $\forall\ \boldsymbol{\zeta}$,* i.e., *given a $\boldsymbol{\zeta}$, $\|\boldsymbol{v}_\theta(\boldsymbol{\zeta},\tau_1)-\boldsymbol{v}_\theta(\boldsymbol{\zeta},\tau_2)\| \leq X_2\|\tau_1-\tau_2\|$ for $\forall\ \tau_1, \tau_2$.*

Assumption 3. $\left\|\boldsymbol{Z}_{t_p}-\boldsymbol{Z}_{t_q}\right\| \leq C\left\|t_p-t_q\right\|, \forall\, p, q \in [0,1]$ *when $\boldsymbol{Z}_{t_p}$ and $\boldsymbol{Z}_{t_q}$ come from the same trajectory.*

Proof. Given a deterministic solver, *e.g.* Euler's method:

$$\boldsymbol{Z}_{t_{i-1}} = \boldsymbol{Z}_{t_i} + (t_{i-1}-t_i)\,\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right). \tag{A.1}$$

The corresponding inversion step of Uni-Inv is denoted by:

$$\widehat{\boldsymbol{Z}}_{t_i} = \widehat{\boldsymbol{Z}}_{t_{i-1}} - (t_{i-1}-t_i)\,\widehat{\boldsymbol{v}}_i, \tag{A.2}$$

where $\widehat{\boldsymbol{v}}_i$ is obtained via Algo. 1 and can be expressed as:

$$\widehat{\boldsymbol{v}}_i = \boldsymbol{v}_\theta\left(\widehat{\boldsymbol{Z}}_{t_{i-1}} - (t_{i-1}-t_i)\,\bar{\boldsymbol{v}}_{i-1}, t_i\right). \tag{A.3}$$

Define the estimation error $\mathcal{E}_i$ as $\mathcal{E}_i = \left\|\boldsymbol{Z}_{t_i} - \widehat{\boldsymbol{Z}}_{t_1}\right\|$. Bringing Eq. A.1 and Eq. A.2 into it, we obtain that:

$$\mathcal{E}_i = (t_{i-1}-t_i)\left\|\widehat{\boldsymbol{v}}_i - \boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right)\right\|. \tag{A.4}$$

Denote $\mathcal{E}_i^1 = \left\|\widehat{\boldsymbol{v}}_i - \boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right)\right\|$, we can bring in Eq. A.3:

$$\mathcal{E}_i^1 = \left\|\boldsymbol{v}_\theta\left(\widehat{\boldsymbol{Z}}_{t_{i-1}} - (t_{i-1}-t_i)\,\bar{\boldsymbol{v}}_{i-1}, t_i\right) - \boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right)\right\|. \tag{A.5}$$

Using the Lipschitz continuity of $\boldsymbol{v}_\theta(\cdot,\tau)$, we have:

$$\mathcal{E}_i^1 \leq X_1\left\|\widehat{\boldsymbol{Z}}_{t_{i-1}} - (t_{i-1}-t_i)\,\bar{\boldsymbol{v}}_{i-1} - \boldsymbol{Z}_{t_i}\right\|. \tag{A.6}$$

Bring in Eq. A.1 for $\boldsymbol{Z}_{t_i}$, there is:

$$\begin{aligned}\mathcal{E}_i^1 &\leq X_1\left\|\left(\widehat{\boldsymbol{Z}}_{t_{i-1}} - \boldsymbol{Z}_{t_{i-1}}\right) + (t_{i-1}-t_i)\left(\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right) - \bar{\boldsymbol{v}}_{i-1}\right)\right\| \\ &\leq X_1\left\|\widehat{\boldsymbol{Z}}_{t_{i-1}} - \boldsymbol{Z}_{t_{i-1}}\right\| + X_1\,(t_{i-1}-t_i)\left\|\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right) - \bar{\boldsymbol{v}}_{i-1}\right\|.\end{aligned} \tag{A.7}$$

The first term is the accumulative error of the previous steps. We denote it as $\mathcal{E}_i^A = \left\|\widehat{\boldsymbol{Z}}_{t_{i-1}} - \boldsymbol{Z}_{t_{i-1}}\right\|$ and it should be neglected for local error analysis. We further denote $\mathcal{E}_i^2 = \left\|\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right) - \bar{\boldsymbol{v}}_{i-1}\right\|$. To analyse this item, we first consider a second-order case, *i.e.*, utilizing an additional function evaluation step to calculate $\bar{\boldsymbol{v}}_{i-1} = \boldsymbol{v}_\theta\left(\widehat{\boldsymbol{Z}}_{t_{i-1}}, t_{i-1}\right)$. Then we have:

$$\mathcal{E}_i^2 = \left\|\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right) - \boldsymbol{v}_\theta\left(\widehat{\boldsymbol{Z}}_{t_{i-1}}, t_{i-1}\right)\right\|. \tag{A.8}$$

Using the Lipschitz continuity of $\boldsymbol{v}_\theta(\cdot,\tau)$ and $\boldsymbol{v}_\theta(\boldsymbol{\zeta},\cdot)$, we have:

$$\begin{aligned}\mathcal{E}_i^2 &= \left\|\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right) - \boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_{i-1}\right) + \boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_{i-1}\right) - \boldsymbol{v}_\theta\left(\widehat{\boldsymbol{Z}}_{t_{i-1}}, t_{i-1}\right)\right\| \\ &\leq \left\|\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right) - \boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_{i-1}\right)\right\| + \left\|\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_{i-1}\right) - \boldsymbol{v}_\theta\left(\widehat{\boldsymbol{Z}}_{t_{i-1}}, t_{i-1}\right)\right\| \\ &\leq X_1\left\|\boldsymbol{Z}_{t_i} - \widehat{\boldsymbol{Z}}_{t_{i-1}}\right\| + X_2\left\|t_i - t_{i-1}\right\|.\end{aligned} \tag{A.9}$$

We denote $\Delta t_i = t_i - t_{i-1}$. Using the Assumption 3, we have:

$$\begin{aligned}\mathcal{E}_i^2 &\leq X_1 \left(\left\| \boldsymbol{Z}_{t_i} - \boldsymbol{Z}_{t_{i-1}} \right\| + \left\| \boldsymbol{Z}_{t_{i-1}} - \widehat{\boldsymbol{Z}}_{t_{i-1}} \right\| \right) + X_2 \Delta t_i \\ &\leq \left(C X_1 + X_2 \right) \Delta t_i + X_1 \mathcal{E}_i^A.\end{aligned} \tag{A.10}$$

Ultimately, the estimation error is as follows:

$$\begin{aligned}\mathcal{E}_i &\leq \Delta t_i \mathcal{E}_i^1 \leq \Delta t_i \left(X_1 \mathcal{E}_i^A + X_1 \Delta t_i \mathcal{E}_i^2 \right) \\ &\leq \Delta t_i \left(X_1 \mathcal{E}_i^A + X_1 \Delta t_i \left(\left(C X_1 + X_2 \right) \Delta t_i + X_1 \mathcal{E}_i^A \right) \right) \\ &= X_1 \left(C X_1 + X_2 \right) \Delta t_i^3 + \left(X_1^2 \Delta t_i^2 + X_1 \Delta t_i \right) \mathcal{E}_i^A.\end{aligned} \tag{A.11}$$

For local error analysis, we neglect the global accumulated error $\mathcal{E}_i^A$, then we have the local error $\mathcal{E}_i^L$:

$$\mathcal{E}_i^L \leq X_1 \left(C X_1 + X_2 \right) \Delta t_i^3 = \mathcal{O}\left(\Delta t_i^3 \right). \tag{A.12}$$

Furthermore, since the step count of the iterative algorithm is $\mathcal{O}\left(1/\Delta t_i\right)$, we can have the global error: $\mathcal{E}^G = \mathcal{O}\left(\max_i \left(\Delta t_i^2 \right) \right)$.

Now, let's go back to the second-order case assumption $\bar{\boldsymbol{v}}_{i-1} = \boldsymbol{v}_\theta \left(\widehat{\boldsymbol{Z}}_{t_{i-1}}, t_{i-1} \right)$ we mentioned earlier. From a practical perspective, Algo. 1 provides an additional function evaluation in the initialization stage, making its first step the standard second-order case. After that, in the ideal case, each $\bar{\boldsymbol{v}}_i$ should converge to $\boldsymbol{v}_i$, and thus the first-order approximation of the algorithm does not significantly affect the error. Theoretically, since that:

$$\begin{aligned}\bar{\boldsymbol{v}}_{i-1} &= \boldsymbol{v}_\theta \left(\bar{\boldsymbol{Z}}_{i-1}, t_{i-1} \right), \\ \bar{\boldsymbol{Z}}_{i-1} &= \widehat{\boldsymbol{Z}}_{i-2} - \left(t_{i-2} - t_{i-1} \right) \bar{\boldsymbol{v}}_{i-2},\end{aligned} \tag{A.13}$$

neglecting the last-step accumulated error, we can derive that

$$\left\| \widehat{\boldsymbol{Z}}_{i-1} - \bar{\boldsymbol{Z}}_{i-1} \right\| \leq \Delta t_{i-1} \left\| \widehat{\boldsymbol{v}}_{i-2} - \bar{\boldsymbol{v}}_{i-2} \right\|. \tag{A.14}$$

Using the Lipschitz continuity of $\boldsymbol{v}_\theta(\cdot, \tau)$, we can get:

$$\left\| \widehat{\boldsymbol{v}}_{i-2} - \bar{\boldsymbol{v}}_{i-2} \right\| \leq X_1 \left\| \widehat{\boldsymbol{Z}}_{i-2} - \bar{\boldsymbol{Z}}_{i-2} \right\|. \tag{A.15}$$

Neglecting the last-step accumulated error of velocity estimation for local error calculation, we note that the one-order approximation brings no change to the conclusion of $\mathcal{E}_i^L = \mathcal{O}\left(\Delta t_i^3 \right)$.

This local error serves the dual processes of inversion and reconstruction, theoretically ensuring the effectiveness of our proposed inversion in practical applications. Moreover, our approach does not utilize the derivative approximation to achieve the result, only expects the velocity function to have good mathematical properties.

B Theoretical Analysis of Uni-Edit

Conducting a mathematical perspective, our proposed Uni-Edit can be compressed into a single arithmetic representation. As shown in Algo. 2, we have the editing result:

$$\begin{aligned}\widetilde{\boldsymbol{Z}}_{\boldsymbol{t_{i-1}}} &= \check{\boldsymbol{Z}}_{t_i} + \left(t_{i-1} - t_i \right) \boldsymbol{v}_i^F \\ &= \widetilde{\boldsymbol{Z}}_{t_i} + \boldsymbol{s}_i + \left(t_{i-1} - t_i \right) \boldsymbol{v}_i^F \\ &= \widetilde{\boldsymbol{Z}}_{t_i} + \omega \left(t_{i-1} - t_i \right) \left(\mathbf{1} + \boldsymbol{m}_i \right) \odot \left(\boldsymbol{v}_i^T - \boldsymbol{v}_i^S \right) \\ &\quad + \left(t_{i-1} - t_i \right) \left(\boldsymbol{m}_i \odot \boldsymbol{v}_i^T + \left(\mathbf{1} - \boldsymbol{m}_i \right) \odot \boldsymbol{v}_i^S \right) \\ &= \widetilde{\boldsymbol{Z}}_{t_i} + \left(t_{i-1} - t_i \right) \boldsymbol{v}_i^*,\end{aligned} \tag{B.16}$$

and the reformed velocity is:

$$\begin{aligned}\boldsymbol{v}_i^* &= \omega \left(\mathbf{1} + \boldsymbol{m}_i \right) \odot \left(\boldsymbol{v}_i^T - \boldsymbol{v}_i^S \right) + \left(\boldsymbol{m}_i \odot \boldsymbol{v}_i^T + \left(\mathbf{1} - \boldsymbol{m}_i \right) \odot \boldsymbol{v}_i^S \right) \\ &= \boldsymbol{v}_i^S + \left(\omega \left(\mathbf{1} + \boldsymbol{m}_i \right) + \boldsymbol{m}_i \right) \odot \left(\boldsymbol{v}_i^T - \boldsymbol{v}_i^S \right),\end{aligned} \tag{B.17}$$

Figure C.1: Illustrations of insufficient editing and background destruction. CLIP_w indicates the whole CLIP similarity.

Table C.1: Component ablation studies of Uni-Edit on PIE-Bench using Stable Diffusion 3. We set step = 15, α = 0.6, and ω = 5.0 in these experiments. "w/o Uni-Inv" means using DDIM Inversion-like Euler inversion to replace Uni-Inv in the editing procedure. "w/o Uni-Edit" indicates using naive delayed injection (using $\boldsymbol{v}_i^T$ after inversion) for editing. "Corr." represents the Correction $\boldsymbol{s}_i$ in Uni-Edit. "Corr. w/o 1+" indicates using $\boldsymbol{m}_i$ as the mask weight of $\boldsymbol{s}_i$ instead of $(1 + \boldsymbol{m}_i)$. "$\boldsymbol{m}_i^{\text{in V.F.}} = \mathbf{1}$" means using $\mathbf{1}$ to replace the mask $\boldsymbol{m}_i$ in Velocity Fusion $\boldsymbol{v}_i^F$ ($\boldsymbol{v}_i^F = \boldsymbol{v}_i^T$), which can be seen as Uni-Edit without Velocity Fusion. "$\boldsymbol{m}_i^{\text{in V.F.}} = \mathbf{0}$" means using $\mathbf{0}$ to replace the mask $\boldsymbol{m}_i$ in $\boldsymbol{v}_i^F$ ($\boldsymbol{v}_i^F = \boldsymbol{v}_i^T$). "$\boldsymbol{m}_i^{\text{in Corr.}} = \mathbf{1}$" indicates using $\mathbf{1}$ to replace the mask $\boldsymbol{m}_i$ in the Correction $\boldsymbol{s}_i$, which can be seen as Uni-Edit without Correction. "$\boldsymbol{m}_i = \mathbf{1}$" claims performing editing without region-adaptive guidance, which is equivalent to using simple classifier-free guidance (CFG).

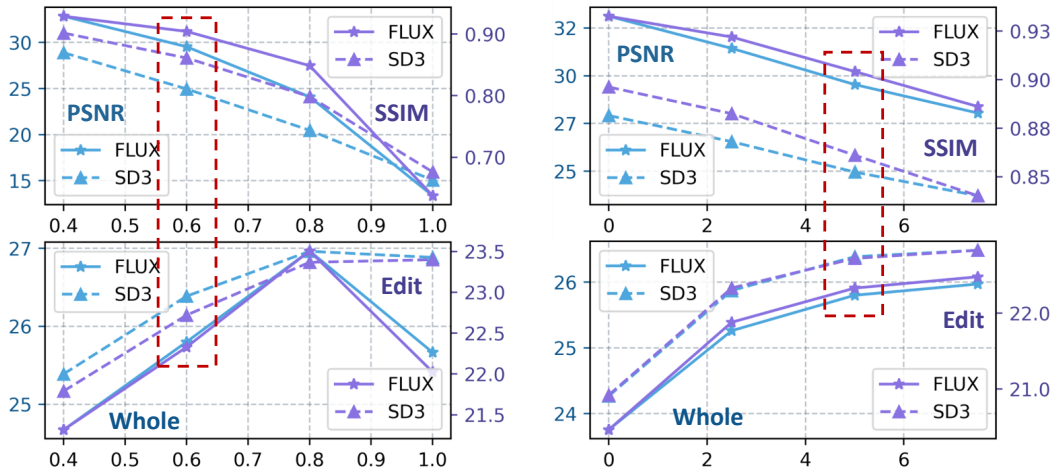
Method	**Structure**	**Background Preservation**		**CLIP Similarity↑**	
	Distance$^{\downarrow}_{10^3}$	**PSNR↑**	**SSIM**$^{\uparrow}_{10^2}$	**Whole**	**Edited**
w/o Uni-Inv	40.87	21.93 (-3.03)	74.90 (-11.21)	25.54 (-0.85)	21.93 (-0.79)
w/o Uni-Edit	9.78	27.92 (+2.96)	89.62 (+3.51)	24.26 (-2.13)	20.92 (-1.80)
w/o Corr.	9.45	28.00 (+3.04)	89.67 (+3.56)	23.78 (-2.61)	20.52 (-2.20)
Inv. $\boldsymbol{v}(\cdot, t_i)$	36.95	23.82 (-1.14)	80.25 (-5.86)	25.99 (-0.40)	22.08 (-0.64)
Corr. w/o 1+	11.33	27.44 (+2.48)	89.31 (+3.20)	25.16 (-1.23)	21.72 (-1.00)
$\boldsymbol{m}_i^{\text{in V.F.}} = \mathbf{1}$	22.92	24.62 (-0.34)	85.50 (-0.61)	26.39 (-0.00)	22.74 (+0.02)
$\boldsymbol{m}_i^{\text{in V.F.}} = \mathbf{0}$	21.78	25.01 (+0.05)	86.13 (+0.02)	26.22 (-0.18)	22.57 (-0.15)
$\boldsymbol{m}_i^{\text{in Corr.}} = \mathbf{1}$	28.98	23.36 (-1.60)	83.13 (-2.98)	26.55 (+0.16)	22.80 (+0.08)
$\boldsymbol{m}_i = \mathbf{1}$	30.48	23.07 (-1.89)	82.52 (-3.59)	26.53 (+0.14)	22.83 (+0.11)
Ours	21.40	24.96	86.11	26.39	22.72

It's interesting that we finally obtain a velocity $\boldsymbol{v}_i^*$ which is very similar to the classifier-free guidance (CFG) (Chung et al., 2022) but with a per-pixel-variant weight instead of a single constant value. Some previous works consider CFG as a predictor-corrector (Song et al., 2020a; Bradley & Nakkiran, 2024). From this perspective, whereas we take a different yet more interpretable approach to conduct a predictor-corrector, and eventually obtain a method with adaptive guidance strength for different regions. In the manuscript, we experimentally validate that the mask obtained from our designed sampling strategy is rationally adaptive to vary with the editing objective and the iteration step. Therefore, our method ensures to achieve per-pixel adaptive guidance strength within the framework of predictor-corrector, which in turn confers effectiveness for text-driven image editing to flow models.

C Ablation Studies of Uni-Edit

C.1 Component Ablations

Preliminary, as shown in Fig. C.2, text-driven image editing tasks pursue a trade-off between editing effect and background preservation. The result we hope for is to preserve non-editing regions while also achieving the editing requirements of the image. Empirically, on PIE-bench (Ju et al., 2024),



(a) Ablation Studies on α (b) Ablation Studies on ω

Figure C.2: **Ablation studies of Uni-Edit** on (a) delay rate α and (b) guidance strength ω. The top indicates the background preservation, while the bottom refers to CLIP scores of editing.

$\text{CLIP}_\text{w} < 25$ always means the editing impact is insignificant (as the blue box in Fig. C.2), while $\text{SSIM} < 84$ usually indicates the background is destructed (as the red circle in Fig. C.2).

We provide ablation studies of the main components of Uni-Edit in Tab. C.1, discussing the impacts of Uni-Inv, Uni-Edit, Correction, Velocity Fusion, and the mask in image editing tasks. Without Uni-Inv (w/o Uni-Inv), the background preservation decreases significantly, indicating that inverted noisy latent, which is capable of accurate reconstruction, is necessary for controllable editing. The results of naive delayed injection (w/o Uni-Edit) show the importance of well-designed guidance for flow-based image editing, which is just as discussed in the manuscript. Meanwhile, the Correction provides guidance targeted to the editing objective, thus unleashing image editing of flow models. When disabling the Correction (w/o Corr.), the result shows almost no editing. Regarding the mask $\boldsymbol{m}_i$, we can demonstrate through relative experiments that it plays an important role in the trade-off between background preservation and editing effect. The mask $\boldsymbol{m}_i$ enhances the correction and editing strength of editing-related regions, while avoiding undesirable influence of these effects on editing-unrelated regions, thereby improving the editing effect and avoiding serious damage to the background. No matter which component disables the mask, it will cause the background preservation to be evidently worse, while the editing effect only has a marginal improvement, showcasing the effectiveness and necessity of the region-adaptive guidance.

To make the ablation studies clearer, we further provide qualitative visual comparisons of the editing results in Fig. C.3. The results in (a) indicate that, without the accurately reconstructable latent provided by Uni-Inv (utilizing a DDIM-Inversion-like inversion method), editing using flow models is likely to crash. Even though inversion using $\boldsymbol{v}(\cdot, t_i)$ has appropriately improved the inversion accuracy, it is still insufficient to support reliable real image editing. Meanwhile, the simple delayed injection without Uni-Edit provides low editing effects, resulting in unchanged images.

Furthermore, (b) shows the results after replacing the fused velocity $\boldsymbol{v}^F$ with $\boldsymbol{v}^T$ or $\boldsymbol{v}^S$. ① illustrates that directly utilizing $\boldsymbol{v}^T$ to move the sample can lead to the background not remaining unchanged. ② demonstrates that if we adopt only $\boldsymbol{v}^S$ as the velocity, even with the correction step, the results can be unnatural (it should have turned into a cup but did not). Additionally, the second row of (b) also indicates that velocity fusion can make the details of the results more reliable (velocity fusion provides the strawberry with a fuller color). Although the velocity itself may not have a strong impact on editing (just like the failure of delayed injection), velocity fusion can still enhance the editing details and provide regional sensitivity, thus making editing more precise and reliable.

Subsequently, (c) provides the editing results of using different inversion velocity functions. The velocity functions here are consistent with Fig. 4. It is evident that without precise inversion like Uni-Inv, achieving success in image editing is difficult, especially for flow models that are suscepti-

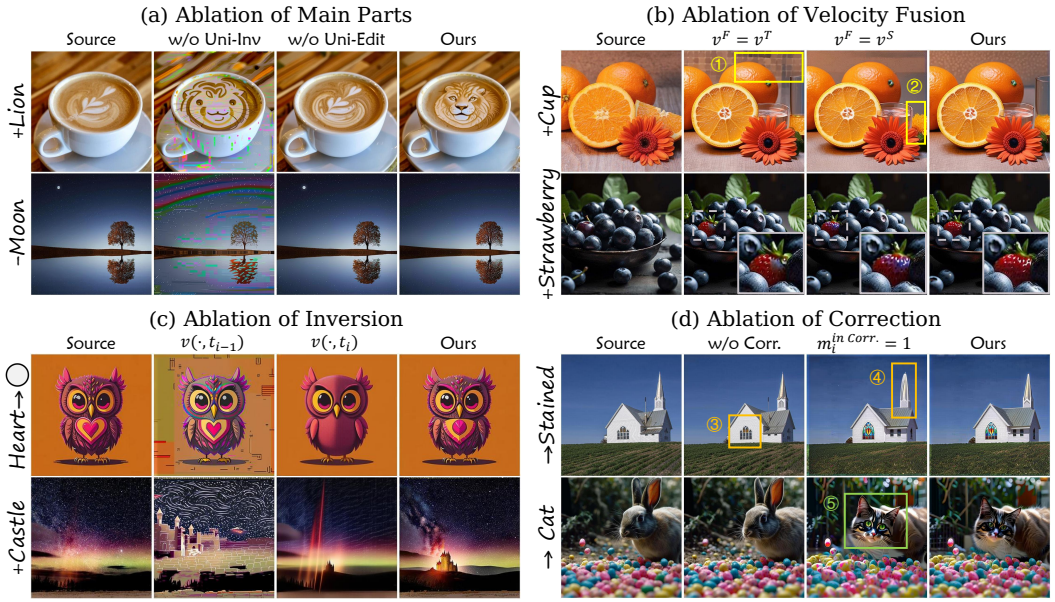


Figure C.3: **Qualitative comparison for ablation studies of Uni-Edit**. (a) Comparison of ablation in the main parts. (b) Comparison of different kinds of editing velocities. (c) Comparison of different kinds of inversion methods. (d) Ablation of correction in Uni-Edit.

ble to cumulative errors. Additionally, (d) presents the visualization of correction's ablation studies. "w/o Corr." means setting the correction step as $\boldsymbol{s}_i = 0$, and $\boldsymbol{m}_i^{\text{in Corr.}} = \mathbf{1}$ denotes turning the correction step into $\boldsymbol{s}_i = \omega(t_{i-1} - t_i)(\mathbf{1}+\mathbf{1}) \odot \boldsymbol{v}_i^-$. The former is similar to the simple delayed injection (instead of using fused velocity), which represents weakening the editing ability of Uni-Edit, while the latter means maintaining the editing ability of Uni-Edit but eliminating region-adaptive guidance. In these cases, ③ shows that without correction the method will significantly reduce editing ability, and ④ indicates that correction without region-adaptive guidance can easily cause changes in editing-irrelevant regions. Not only that, as shown in ⑤, correction without regional restrictions can also easily cause overexposure, which is similar to large classifier-free guidance (CFG). These visualizations of ablation results further demonstrate the roles and significance of the components in our proposed method.

C.2 Hyper-parameter Selection

Additionally, we present the ablation studies of our proposed Uni-Edit with step = 15 for various hyper-parameters in Fig. C.2. The results demonstrate that different hyper-parameters bring rational skews to the trade-off between background preservation and editing effectiveness. Nevertheless, our approach improves the overall level of the trade-off, making it effortless to bring benefits to both aspects.

D Uni-Inv on Different Generation Methods

D.1 Heun Method based Uni-Inv

To validate the transferability and effectiveness of our method on different samplers, we reimplement our Uni-Inv based on the Heun method. To be specific, since the Heun method is formulated as:

$$\boldsymbol{Z}_{t_{i-1}} = \boldsymbol{Z}_{t_i} + (t_{i-1} - t_i)\frac{\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right) + \boldsymbol{v}_\theta\left(\boldsymbol{v}_\theta\left(\boldsymbol{Z}_{t_i}, t_i\right), t_i\right)}{2}, \tag{D.18}$$

we directly reform the velocity function as:

$$\boldsymbol{v}_\theta^H\left(\boldsymbol{\zeta}, \tau\right) = \frac{\boldsymbol{v}_\theta\left(\boldsymbol{\zeta}, \tau\right) + \boldsymbol{v}_\theta\left(\boldsymbol{v}_\theta\left(\boldsymbol{\zeta}, \tau\right), \tau\right)}{2}, \tag{D.19}$$

Table D.2: **Quantitative results for inversion and reconstruction** of our Uni-Inv based on Heun method with Flow models and DDIM with Diffusion models. We set the step to 50 for SDXL (Podell et al., 2023), 25 for SD3 (Esser et al., 2024), and 15 for FLUX models to conduct the experiments. The best results are **bolded**.

Method	Model	Unconditional				Conditional			
		$\text{MSE}^{\downarrow}_{10^3}$	PSNR$\uparrow$	$\text{SSIM}^{\uparrow}_{10^2}$	$\text{LPIPS}^{\downarrow}_{10^2}$	$\text{MSE}^{\downarrow}_{10^3}$	PSNR$\uparrow$	$\text{SSIM}^{\uparrow}_{10^2}$	$\text{LPIPS}^{\downarrow}_{10^2}$
DDIM	SDXL	8.99	22.19	75.57	12.76	7.35	23.21	77.73	10.20
Ours (DDIM)		**6.32**	**24.18**	**79.05**	**8.60**	**5.17**	**25.04**	**80.63**	**6.95**
Heun	SD3	25.34	16.98	67.25	26.63	26.32	16.89	64.14	27.70
Ours (Heun)		**20.23**	**20.10**	**76.38**	**15.76**	**12.75**	**22.31**	**79.62**	**12.43**
Heun	FLUX	83.04	11.77	42.10	39.96	76.79	12.17	40.17	41.18
Ours (Heun)		**57.39**	**13.63**	**57.45**	**26.95**	**32.35**	**16.79**	**67.33**	**21.25**

Table D.3: Inversion comparison between iterative inversion methods and Uni-Inv on the first 500 images of CC3M.

Method	Model	Unconditional			Conditional			Steps	NFE
		PSNR$\uparrow$	$\text{SSIM}^{\uparrow}_{10^2}$	$\text{LPIPS}^{\downarrow}_{10^2}$	PSNR$\uparrow$	$\text{SSIM}^{\uparrow}_{10^2}$	$\text{LPIPS}^{\downarrow}_{10^2}$		
ReNoise	Diffusion	24.14	78.71	11.99	24.29	78.95	11.66	50	150
GNRI	Diffusion	23.90	78.07	8.68	23.88	78.03	8.79	50	150
Ours	Diffusion	**24.41**	**79.67**	**7.96**	**25.24**	**81.08**	**6.31**	50	101
ReNoise	SDXL-Turbo	17.59	57.23	28.43	16.81	55.10	29.27	4	16
GNRI	SDXL-Turbo	13.46	49.43	39.82	13.20	48.56	38.68	4	12
Ours	SDXL-Turbo	**20.08**	**71.15**	**14.32**	**20.63**	**71.86**	**13.14**	4	9

then using $\boldsymbol{v}_\theta^H$ to replace the original velocity function in Algo. 1. As shown in Tab. D.2, our approach improves the reconstruction accuracy of the Heun method across the board. This demonstrates the flexibility and adaptability of Uni-Inv to different samplers and reflects its effectiveness.

D.2 DDIM BASED UNI-INV

DDIM (Song et al., 2020a) provides an efficient sampling method for the stochastic-differential-equation-based diffusion models and allows access to the sampling strategy with the form of ordinary differential equations. Benefiting from this, we are able to migrate Uni-Inv to diffusion models by simply treating the predicted initial noise as a velocity, utilizing the cached last-step predicted noise to push forward the current samples to the next timestep, thus performing our inversion. We evaluate the above strategy on SDXL (`RealVisXL_V4.0`) (Podell et al., 2023). The results are shown in Tab. D.2. Though DDIM has already demonstrated strong feasibility in numerous applications, our approach can still take it to the next level and bring about an overall improvement in reconstruction accuracy. It also indicates that our work does not just face a particular methodology. We expect to build approaches that can continuously provide insights into the developing trend of generative models.

D.3 COMPARISON BETWEEN ITERATIVE INVERSION METHODS AND UNI-INV

We further compare iterative inversion methods (Garibi et al., 2024; Samuel et al., 2025) with our proposed Uni-Inv on inversion and reconstruction experiments. As there are no flow-based implementations of these methods, we compare inversion on diffusion using official code and settings in D.3. For fairness, we set the optimization steps on SDXL-Turbo of ReNoise (Garibi et al., 2024) to 2 as GNRI (Samuel et al., 2025) does. We adopt the same experimental settings as the manuscript, except for the sampling steps (50 for diffusion models and 4 for SDXL-Turbo). The experiments on

the first 500 samples of the CC3M (Sharma et al., 2018) dataset further provide Uni-Inv's superiority compared with the mentioned iterative inversion methods.

E Uni-Edit on Diffusion Models

Table E.4: **Text-driven image editing comparison** on PIE-Bench (Ju et al., 2024) based on Diffusion models. We evaluate our proposed Uni-Edit using SDXL (Podell et al., 2023). We keep the same hyper-parameter setting with our main experiments (*i.e.*, $\alpha = 0.6$ and $\omega = 5$), and adopt 50 and 15 as steps. Besides tuning-based methods are marked in gray, the best and second best results are **bolded** and underlined, respectively.

Method	Model	Struc.	BG Preservation				CLIP Sim.↑		Steps	NFE
		Dist.$^{\downarrow}_{10^3}$	PSNR↑	LPIPS$^{\downarrow}_{10^3}$	MSE$^{\downarrow}_{10^4}$	SSIM$^{\uparrow}_{10^2}$	Whole	Edited		
Null-Text Inv	Diff.	13.44	27.03	60.67	35.86	84.11	24.75	21.86	50	-
ReNoise	Diff.	-	27.11	49.25	31.23	72.30	23.98	21.26	50	-
P2P	Diff.	69.43	17.87	208.80	219.88	71.14	25.01	22.44	50	100
P2P-Zero	Diff.	61.68	20.44	172.22	144.12	74.67	22.80	20.54	50	100
PnP	Diff.	28.22	22.28	113.46	83.64	79.05	25.41	22.55	50	100
PnP-Inv.	Diff.	24.29	22.46	106.06	80.45	79.68	25.41	22.62	50	100
EditFriendly	Diff.	-	24.55	91.88	95.58	81.57	23.97	21.03	50	100
MasaCtrl	Diff.	28.38	22.17	106.62	86.97	79.67	23.96	21.16	50	100
InfEdit	Diff.	**13.78**	**28.51**	**47.58**	**32.09**	**85.66**	25.03	22.22	12	72
Ours	Diff.	15.59	25.64	78.83	43.05	83.42	**26.33**	**22.78**	15	28

Similar to Uni-Inv, since our proposed Uni-Edit is completely sample-based and model-agnostic, it is also capable of migrating to diffusion models effortlessly. We adopt the SDXL (`RealVisXL_V4.0`) (Podell et al., 2023) as our base model, conducting evaluation experiments on PIE-Bench (Ju et al., 2024). The results are shown in Tab. E.4. Here we enumerate the previous SOTA of diffusion-based approaches, wherein compared to the manuscript, we additionally present the results of tuning-based inversion (Mokady et al., 2023; Garibi et al., 2024) applied to editing. In contrast to these approaches, the CLIP similarity metrics exemplify our proposed Uni-Edit's capability to drive the diffusion-based editing to new heights and maintain highly competitive background preservation. Meanwhile, we significantly improve the editing efficiency by reducing NFE extremely compared to previous diffusion-based approaches. These experiments strongly demonstrate the effectiveness, adaptability, and generalizability of our proposed approaches, providing new insights into image inversion and editing in the era of flow models.

Furthermore, there are still many training-based methods (Wu et al., 2024; Brooks et al., 2023; Shi et al., 2024; Wei et al., 2024) to learn how to reasonably edit images from the provided training data. Most of them focus on conducting flexible editing through user-provided instructions. Such ideas are very practical and effective. Nonetheless, before embarking on this kind of approach, it is crucial to clearly learn about the properties of the base generation methods. This is also the main concentration of this paper.

F Diverse Applications

In addition to general text-driven image editing, the interpretable design of our method enables a wide range of applications. Fig. F.4 showcases its use for sketch-to-image (1st line) and stroke-to-image (2nd line) tasks (Yu et al., 2015; Chowdhury et al., 2022). For these applications, we set $\alpha = 0.8$ to enhance the editing effect. By exploiting the binary nature of sketches and fixing $\boldsymbol{m}_i$ as their grayscale value, we achieve more robust results for sketch-to-image tasks. Moreover, thanks to the advanced flow matching-based video generation model Wan (WanTeam et al., 2025), we further test Uni-Edit on video editing tasks. We directly consider the latent containing the temporal dimension as $\boldsymbol{Z}$, and apply Uni-Edit to Wan's sampling process without any modification. Setting $\alpha = 0.8$ and

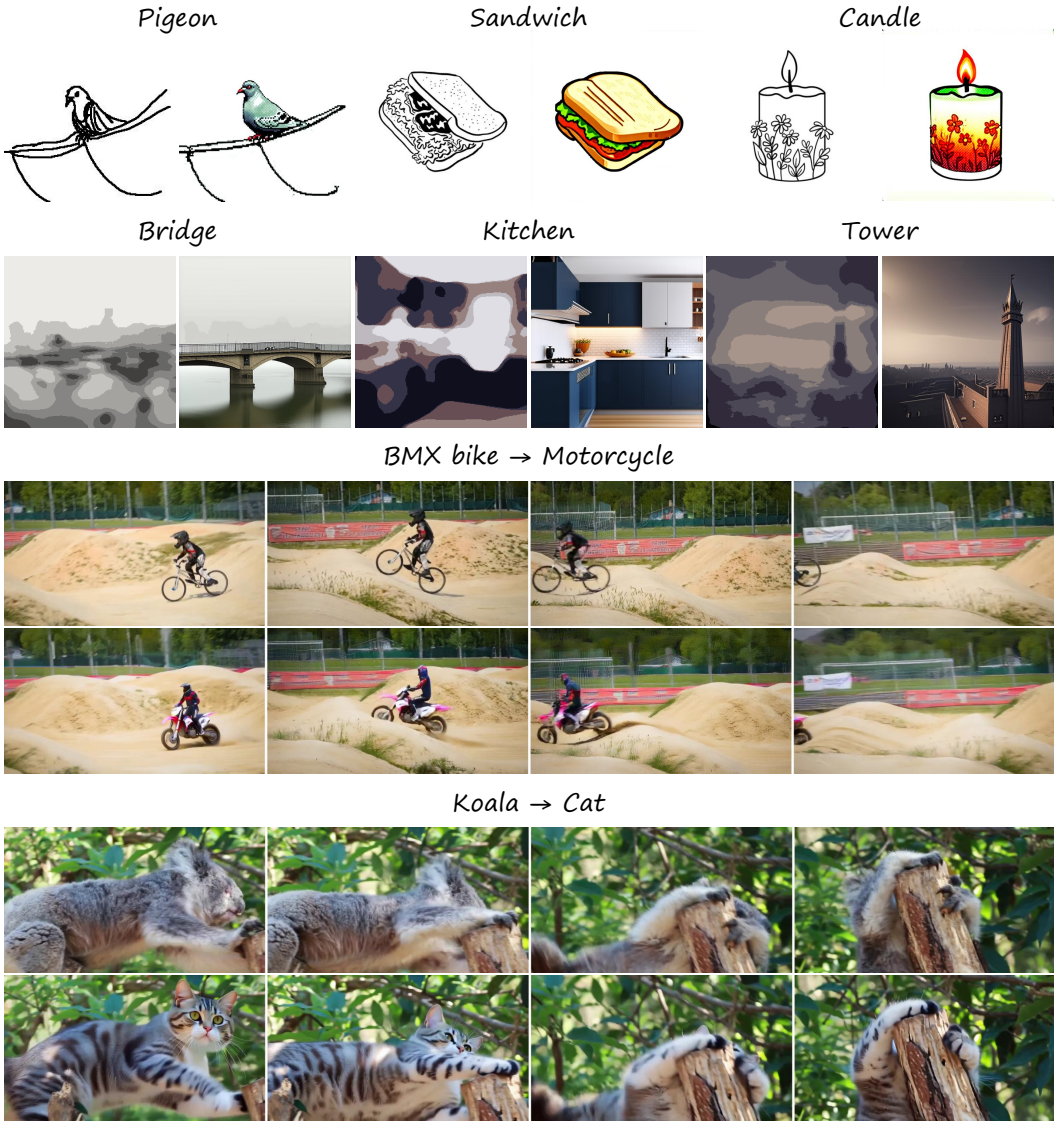


Figure F.4: **Diverse application of Uni-Edit**. The top is sketch to image, and the bottom is stroke to image. The left of a image pair is the source image, and the right is the editing result.

$N = 25$, we achieve reliable editing results (3rd and 4th parts) using `Wan2.1-T2V-1.3B` model. These results further highlight the generalizability and effectiveness of our approach.

G Application Utilizing Diversified Plugins

Since our proposed method is model-agnostic, various plugins that can insert flow models can be applied to provide different editing conditions or to achieve specific editing objectives. These plugins can generally provide ***new conditions*** or help enhance ***controllability***, enabling image editing to meet different specific needs. Therefore, in the era of rapid development of generative models represented by flow models, our method can stably and continuously integrate into stronger models or more complex tasks.

G.1 Introducing of New Conditions

Many previous works achieve personalized generation based on images by inserting image features into prompt embeddings or attentions, among which IP-Adapter (Ye et al., 2023) is one of the most

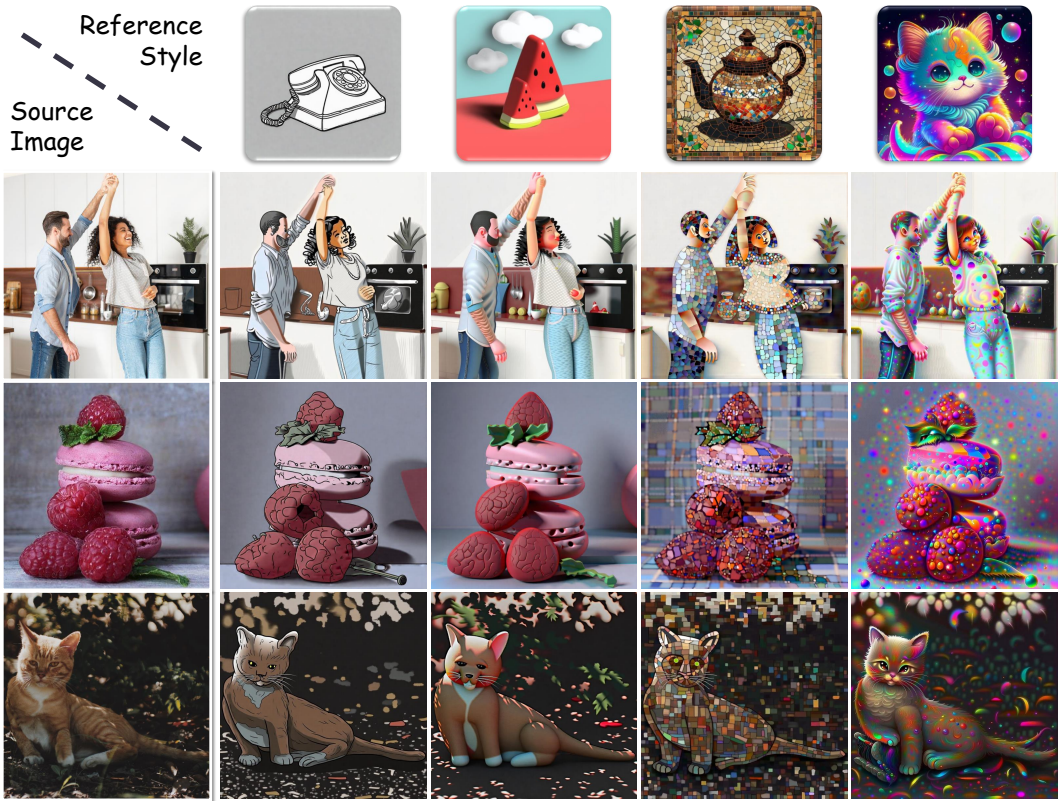


Figure G.5: **Applications of Uni-Edit utilizing IP-Adapter (Ye et al., 2023; Team, 2024) for reference-based style transfer**. The first column is the source image, and the first row is the reference style image.

representative approaches. Taking inspiration from this, we first attempted to apply an IP-Adapter that facilitates style transformation (Team, 2024) to our pipeline. During the editing process of Uni-Edit, we load `InstantX/FLUX.1-dev-IP-Adapter` into the FLUX model and differentiate between source and target conditions by distinguishing among different image inputs.

Specifically, we first employ the original Uni-Inv conditioned on null text, and then modify the $\boldsymbol{v}_i^S$ and $\boldsymbol{v}_i^T$ in Uni-Edit. After adopting the IP-Adapter, the velocity function becomes $\boldsymbol{v} = \boldsymbol{v}_\theta(\boldsymbol{Z}_t, t \mid \boldsymbol{c}_{\text{txt}}, \boldsymbol{c}_{\text{img}})$, where $\boldsymbol{c}_{\text{img}}$ denote the input image of the IP-Adapter. Subsequently, in order to make the image editing focus on style transfer, we keep the text conditions of $\boldsymbol{v}_i^S$ and $\boldsymbol{v}_i^T$ consistent with $\boldsymbol{c}_{\text{txt}}^S$, without introducing any changes to the content:

$$\boldsymbol{v}_i^S = \boldsymbol{v}_\theta(\widetilde{\boldsymbol{Z}}_{t_i}, t_i \mid \boldsymbol{c}_{\text{txt}}^S, \boldsymbol{c}_{\text{img}}^S), \quad \boldsymbol{v}_i^T = \boldsymbol{v}_\theta(\widetilde{\boldsymbol{Z}}_{t_i}, t_i \mid \boldsymbol{c}_{\text{txt}}^S, \boldsymbol{c}_{\text{img}}^T). \tag{G.20}$$

Fig. G.5 shows the results of style transfer using our proposed Uni-Edit on IP-Adapter-injected FLUX model. We set $\alpha = 0.6, \omega = 5.0, N = 15$ for Uni-Edit here. We adopt the source image as $\boldsymbol{c}_{\text{img}}^S$ and the reference style image as $\boldsymbol{c}_{\text{img}}^T$. It can be clearly observed that the editing results accurately capture the style of the reference style image (such as 3D rendering style, colorful style, etc.). At the same time, our method does not cause excessive damage to the source image, allowing the edited results to maintain both the targeted style features and the original content.

In addition, under the same paradigm, we have also adopted another type of IP-Adapter, InstantCharacter (Tao et al., 2025), which has the ability to customize the characters in the generated images. We utilize the source image as $\boldsymbol{c}_{\text{img}}^S$ and the reference character image as $\boldsymbol{c}_{\text{img}}^T$ to achieve image character editing, and set $\alpha = 0.7, \omega = 3.0, N = 30$ here. The results are shown in Fig. G.6, which showcases the abilities of our method using InstantCharacter for effective face editing.

These experiments not only demonstrate the strong flexibility and diverse application scenarios of our proposed method, but also illustrate the promising scalability of such a sampling based strategy.

Figure G.6: **Applications of Uni-Edit utilizing InstantCharacter (Tao et al., 2025) for face editing**. In each group, the upper left is the source image, the lower left is the reference character image, and the right is the editing result.

G.2 Enhancement of Controllability

Previous work has also proposed many modules that inject additional control conditions, among which ControlNet (Zhang et al., 2023) is the most representative. We can introduce effective control during the editing process of Uni-Edit by treating these injected conditions as part of the velocity function, *i.e.*:

$$\boldsymbol{v}'_\theta(\widetilde{\boldsymbol{Z}}_t, t \mid \boldsymbol{c}_{\text{txt}}) = \boldsymbol{v}_\theta(\widetilde{\boldsymbol{Z}}_t, t \mid \boldsymbol{c}_{\text{txt}}, \boldsymbol{c}_{\text{ctrl}}), \tag{G.21}$$

where $\boldsymbol{c}_{\text{ctrl}}$ denotes the input condition of the ControlNet. By replacing $\boldsymbol{v}_\theta$ in Alg. 2 with $\boldsymbol{v}'_\theta$, it can be ensured that the control conditions are preserved in the image during the inversion and editing process.

Fig. G.7 shows the editing results with enhanced controllability. We utilize Stable Diffusion 3 with Canny-conditioned ControlNet (`InstantX/SD3-Controlnet-Canny`) as the base model for our Uni-Inv and Uni-Edit, and set $\alpha = 0.9, \omega = 5.0, N = 30$. These images are from the GTAV dataset (Richter et al., 2016), and the Canny edges used for control are the Canny edges of the segmentation labels of these images. We utilize null text as the source prompts and the word describing environment ("Snowy", "Rainy", "Foggy", and "Night") as the target prompts in these experiments. The results indicate that after introducing the control of ControlNet, Uni-Edit exhibits strong semantic information retention abilities and also achieves significant editing of environmental features in the image. This application can be used in autonomous driving scenarios or world model building processes to provide more diverse while reliable data for training semantic segmentation and detection recognition models.

H Additional Qualitative Comparison

H.1 Uni-Inv

We represent more qualitative comparison results of our Uni-Inv and recent flow-based approaches (Wang et al., 2024b; Deng et al., 2024) in Fig. H.8 and Fig. H.9. These figures contain a wide variety of image samples, including landscape photographs, object photographs, human-centered daily photographs, photographs in extreme lighting, group photographs of large numbers of people,

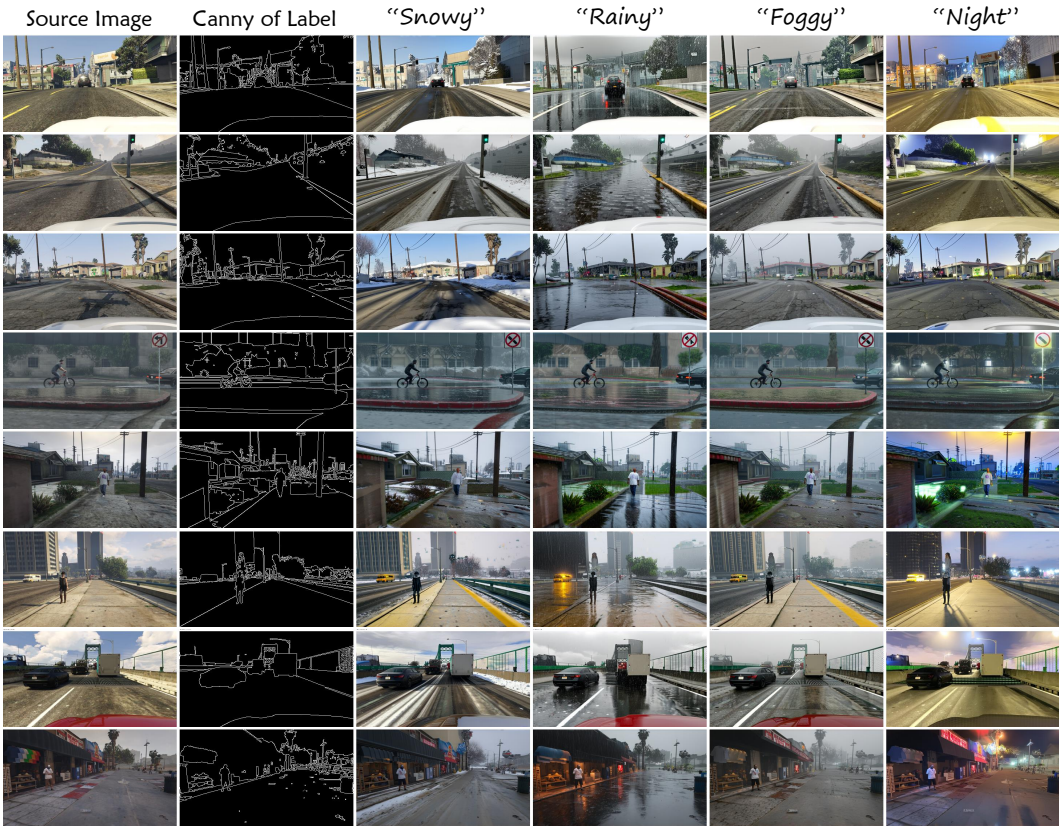


Figure G.7: **Application of Uni-Edit for reliable environment style transformation for autonomous driving tasks using ControlNet (Zhang et al., 2023).** The first column is the source image, and the second column is the reference canny image, which is obtained from the ground truth segmentation label of the source image. The first row provides editing prompts of such editing tasks (only one discription word is enough).

black and white photographs, posters, pencil drawings, oil paintings, etc. of varying resolutions. Our method well maintains the overall image color (last line of Fig. H.8), texture style (6th line of Fig. H.8), content details including text (8th and 9th lines of Fig. H.9) during inversion & reconstruction, achieving consistent superiority in both conditional and unconditional settings.

H.2 UNI-EDIT

We further perform additional qualitative comparisons with existing state-of-the-art methods (Hertz et al., 2022; Tumanyan et al., 2023; Ju et al., 2024; Huberman-Spiegelglas et al., 2024; Cao et al., 2023; Xu et al., 2024a; Rout et al., 2024; Wang et al., 2024b; Deng et al., 2024) on text-driven image editing as shown in Fig. H.10, Fig. H.11, and Fig. H.12. We extensively compared the different approaches under conditions of editing categories, materials, properties, motions, backgrounds, and types of adding or removing items or concepts, as well as stylization.

First, in the task of regional editing, our method demonstrates significant local perception and background preservation capabilities. It is worth noting that our approach is model-agnostic, *i.e.* it does not require the involvement of the attention mechanism. This leads to more oriented regional editing. For example, attention-based methods such as PnP (Tumanyan et al., 2023) often impose attributes that need to be used for regional editing on irrelevant regions (4th line of Fig. H.11). Due to these methods disassembling prompts and attempting to utilize a single token about the editing to exert guidance, inappropriate semantic understanding comes since the wholeness of prompts is destroyed. On the contrary, our approach is model-agnostic, thus better capitalizing on the text comprehension capabilities learned from the model.

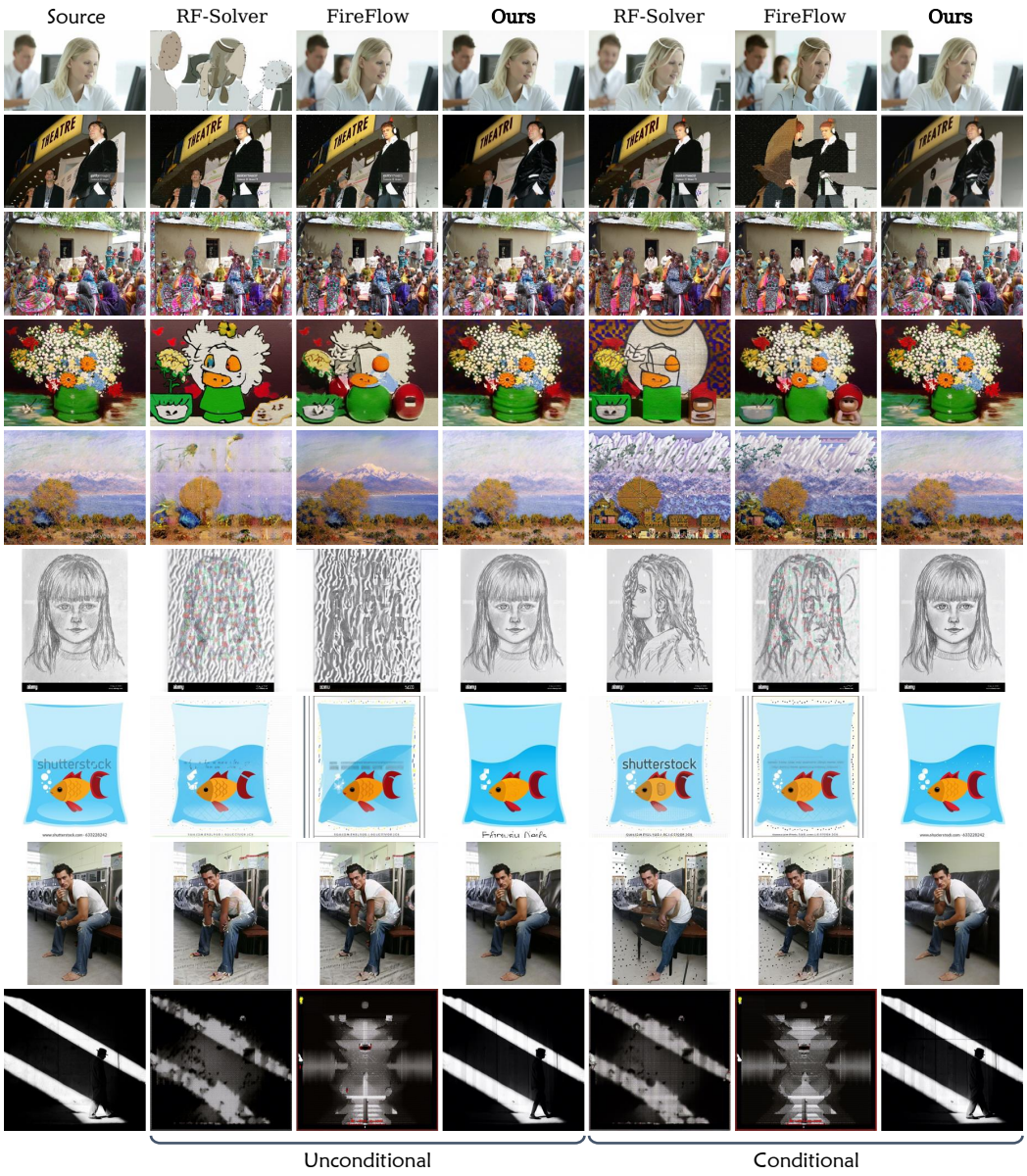


Figure H.8: **Additional qualitative comparison on inversion & reconstruction** on the Conceptual Captions validation dataset (Sharma et al., 2018).

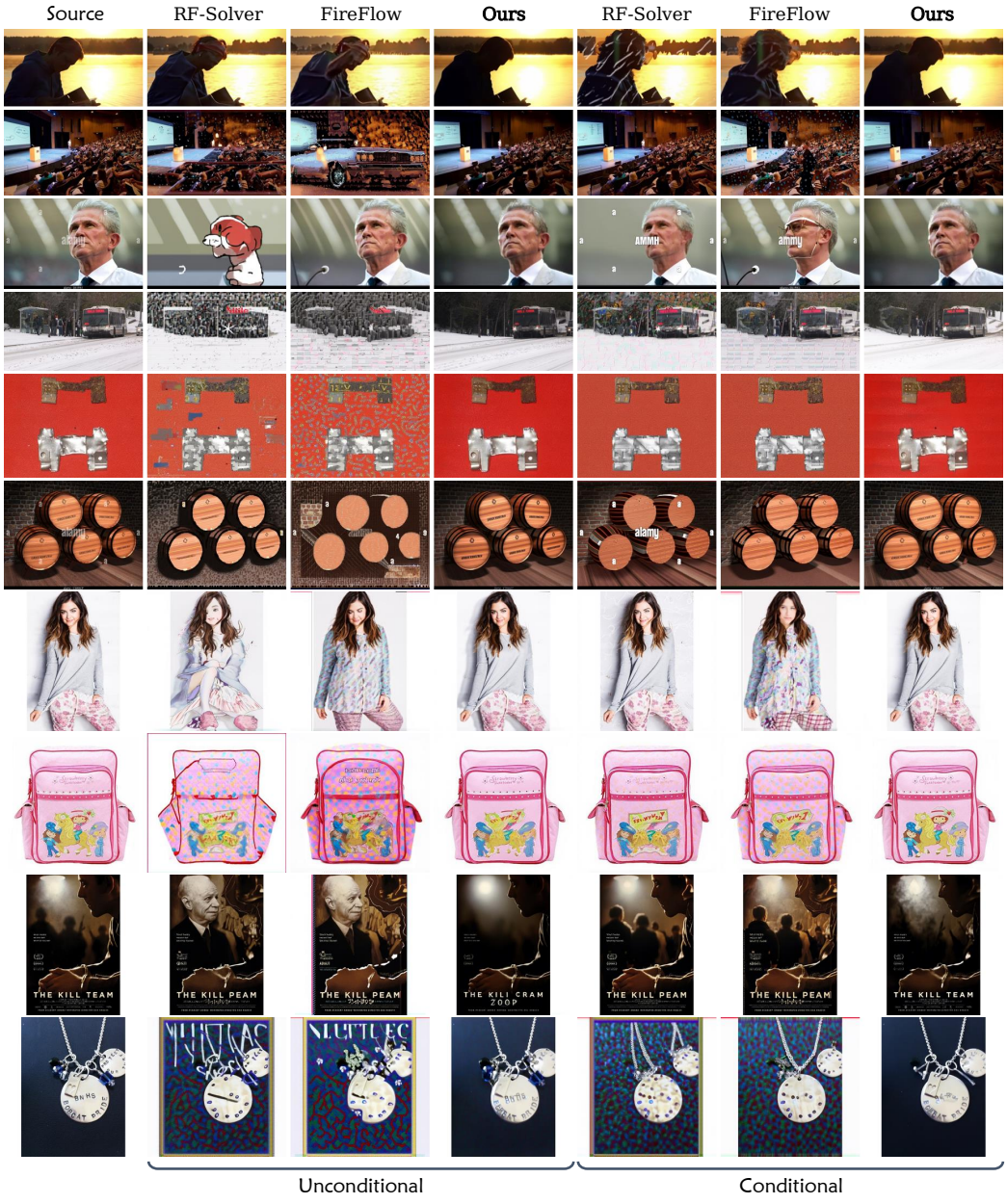


Figure H.9: **Additional qualitative comparison on inversion & reconstruction** on the Conceptual Captions validation dataset (Sharma et al., 2018).

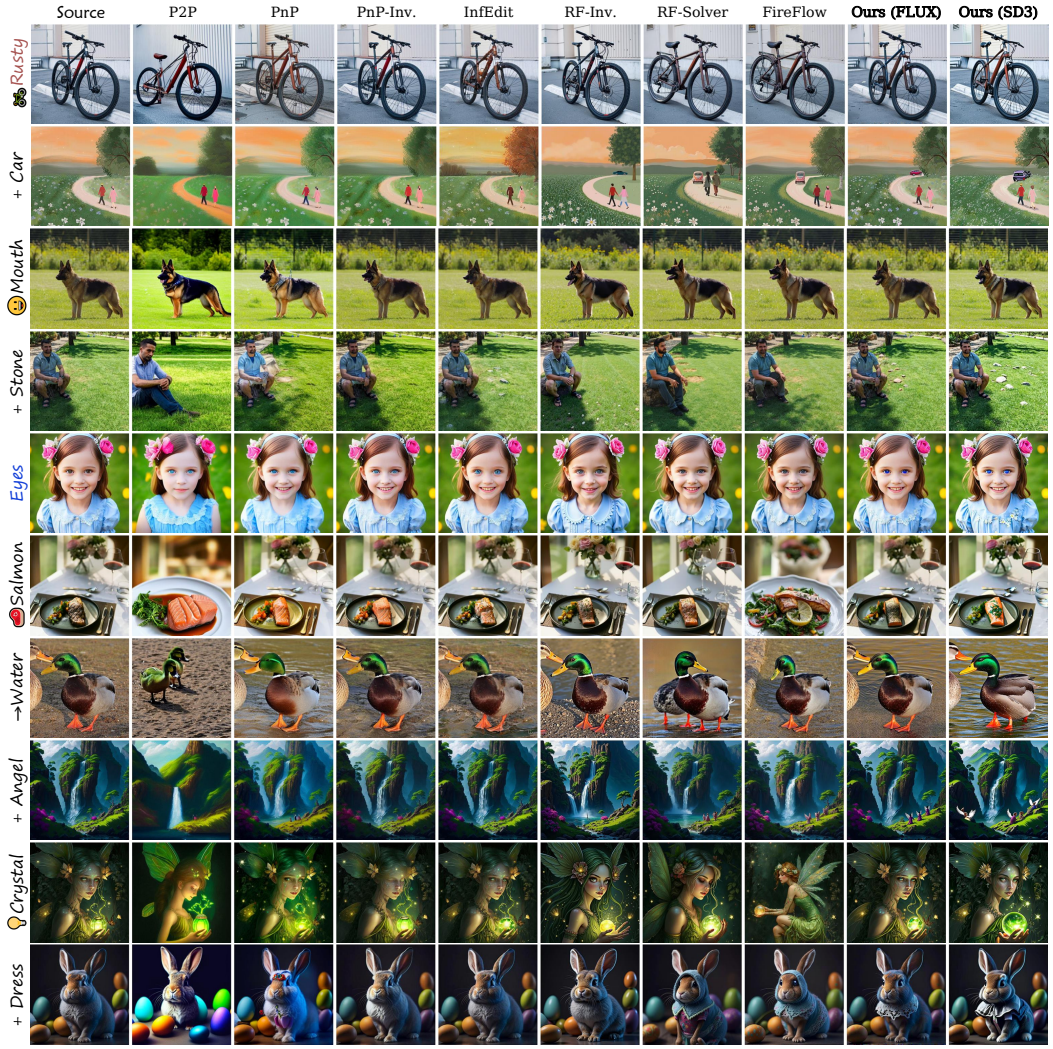


Figure H.10: **Additional qualitative comparison on image editing** on PIE-Bench (Ju et al., 2024).

Subsequently, compared to the same sampling-based kind of approaches (*i.e.*, RF-Inversion (Rout et al., 2024), RF-Solver (Wang et al., 2024b), and FireFlow (Deng et al., 2024)), our method demonstrates significant advantages in terms of image structure and background preservation while maintaining robust editing (1st and 9th lines of Fig. H.10, 3rd and last lines of Fig. H.11, 3rd and 7th lines of Fig. H.12). On the one hand, this is due to our proposed Uni-Inv theoretically ensuring a small local error in the inversion process that can support accurate reconstruction. On the other hand, our deep exploration and re-empowerment of delayed injection make it easy for our proposed Uni-Edit to strike a satisfying balance between editing and the preservation of editing-irrelevant concepts.

I Additional Results on Editing Tasks

I.1 Image Editing

Fig. I.13 and Fig. I.14 represent additional qualitative results of our proposed Uni-Edit on image editing tasks. These results indicate that when it comes to diverse targets and diverse image domains, our approach still remains very effective. It is worth noting that each edit in Fig. I.13 contains multiple different editing objectives (*e.g.*, changing the time, removing the crowd, and adjusting the lighting). Our approach is able to simultaneously achieve these various targets in a single round, using only the original prompt and the target prompt as guidance. Benefiting from the sampling-

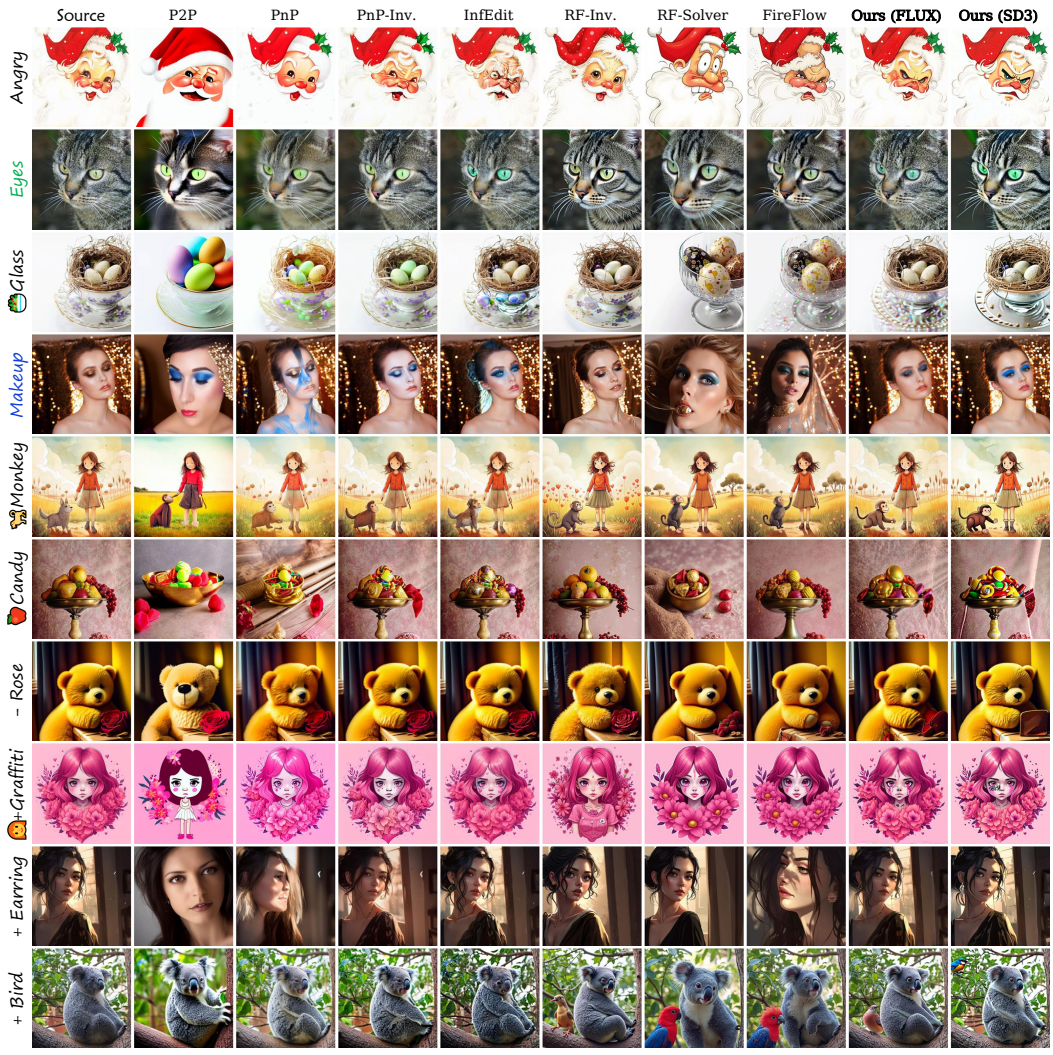


Figure H.11: **Additional qualitative comparison on image editing** on PIE-Bench (Ju et al., 2024).

Figure H.12: **Additional qualitative comparison on image editing** on PIE-Bench (Ju et al., 2024).

based design, our approach is able to capture diverse objectives at once through the relationship between the latents obtained from different conditions. Compared to methods that rely on cross attention manipulations, this design is simpler, more robust, and less likely to cause confusion.

I.2 Video Editing

Moreover, we directly adopt Uni-Edit to conduct video editing tasks using the flow matching-based video generation model Wan (WanTeam et al., 2025). Qualitative results are shown in Fig. I.15. Since our method is model-agnostic, we can achieve reliable video editing results without additional design or complex parameterization. It is further strong evidence of our approach's generalizability.

J Limitations and Future Works

The core issue plaguing us now is that our Uni-Edit is designed for image-text pair inputs. It is not capable of accepting more than one image as the condition. This results in no direct way for us to contribute to the personalization generation problems. In the future, we would like to develop editing methods that are more general and oriented to more diverse tasks. The accurate inversion of Uni-Inv helps to capture image information. With this facilitation, we hope to develop sampling-based editing strategies capable of injecting image conditions based on our re-enabled delayed injection framework. We leave it as an interesting future work.

K LLM Usage Statement

In this paper, LLMs were not used for polishing writing, discovery and retrieval, research ideation, and other aspects. All paper writing, scientific content, and interpretations are the authors' own.

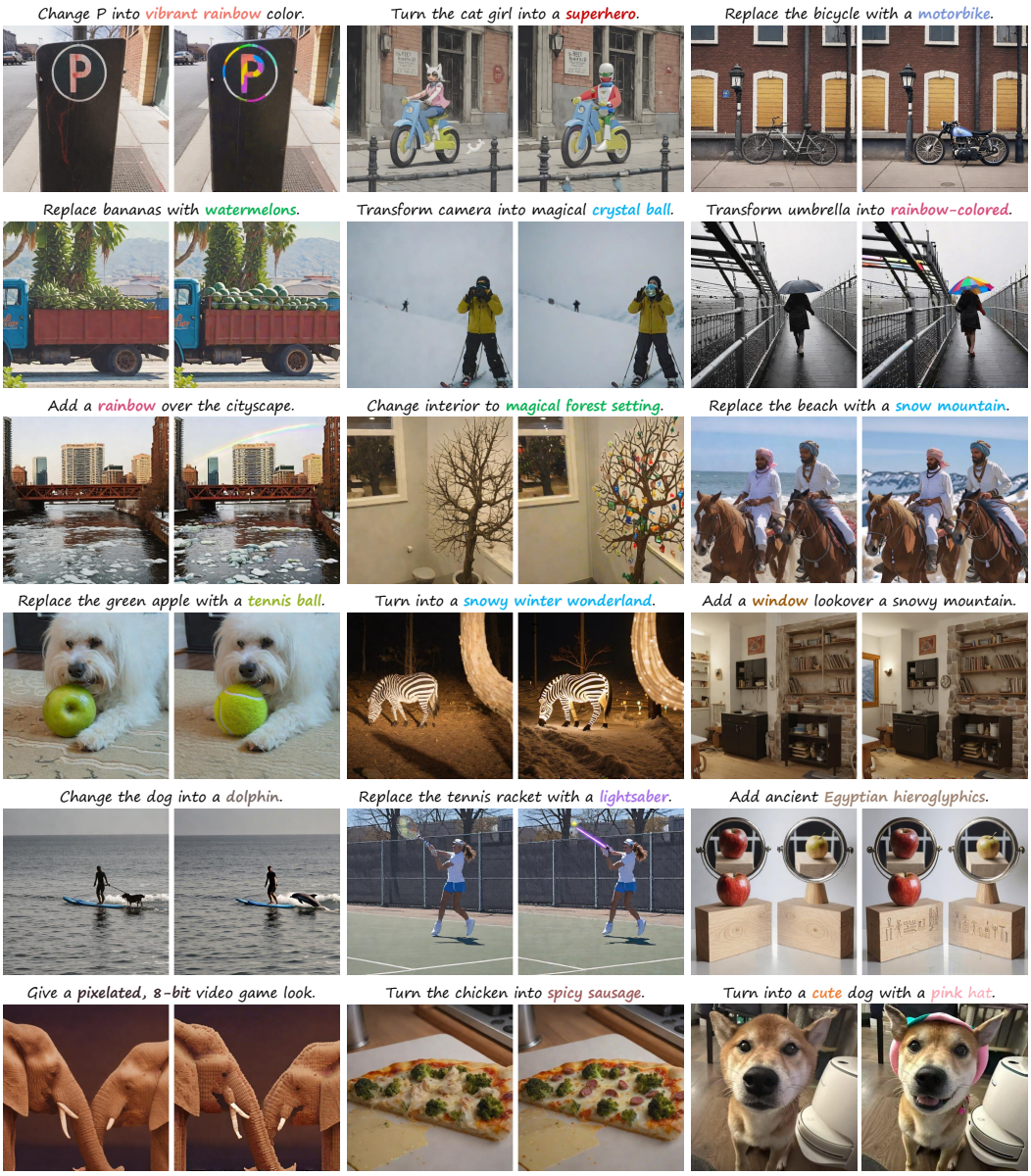


Figure I.13: **Additional qualitative results on image editing** on UltraEdit dataset (Zhao et al., 2024) and wild images. In each image pair, the left is the original image and the right is the result of our editing. The text captions at the top of images are descriptions of the editing objectives and are not the input to the model. We still maintain the paradigm of using the original prompt and the target prompt as conditions. These images are obtained by FLUX using Uni-Edit with $\alpha = 0.6$, $\omega = 5$ and step = 15 which is consistent with the main experiments.

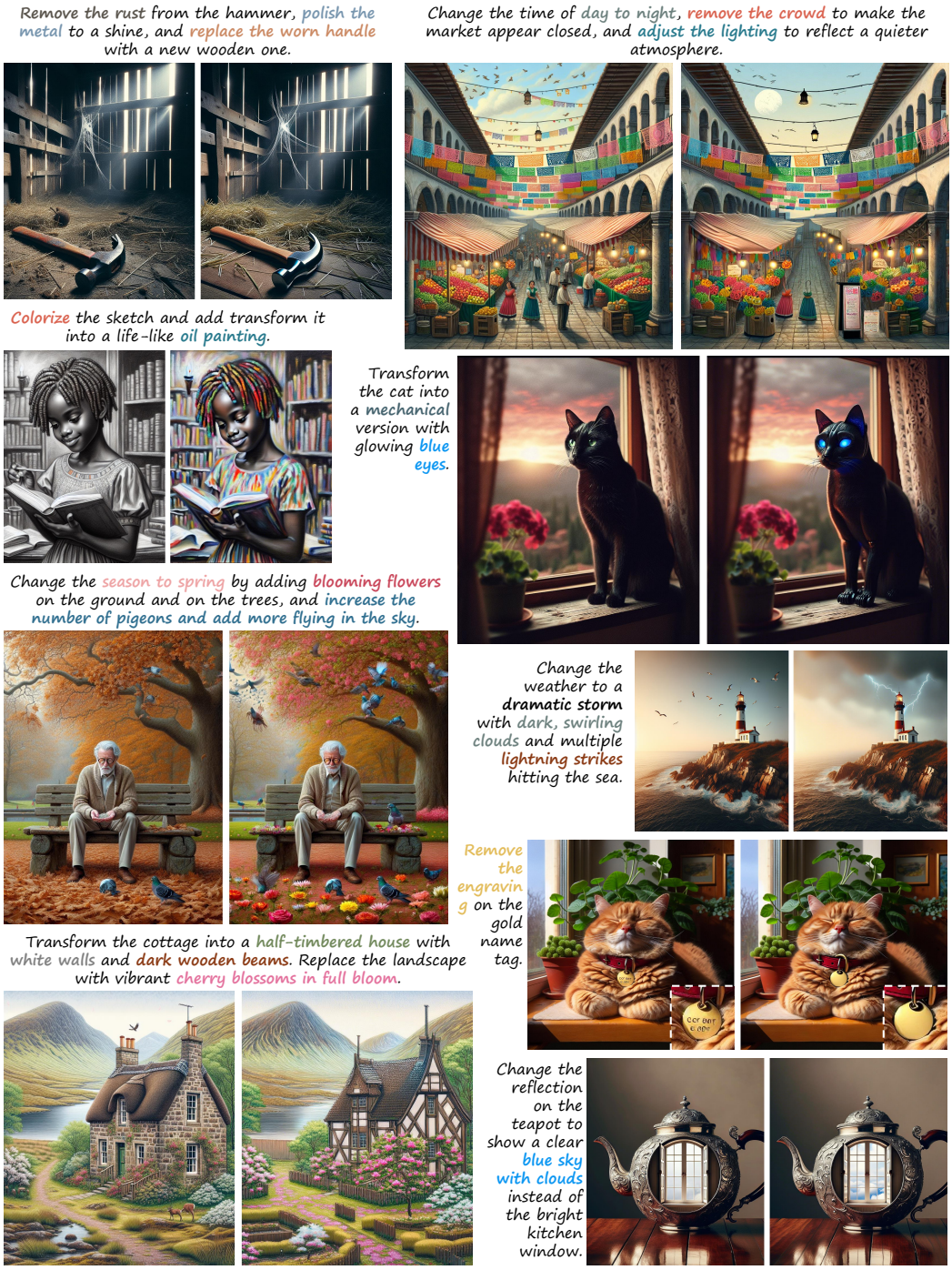


Figure I.14: **Additional qualitative results on image editing** on HQ-Edit dataset (Hui et al., 2024). The visualization setup is the same as Fig. I.13.

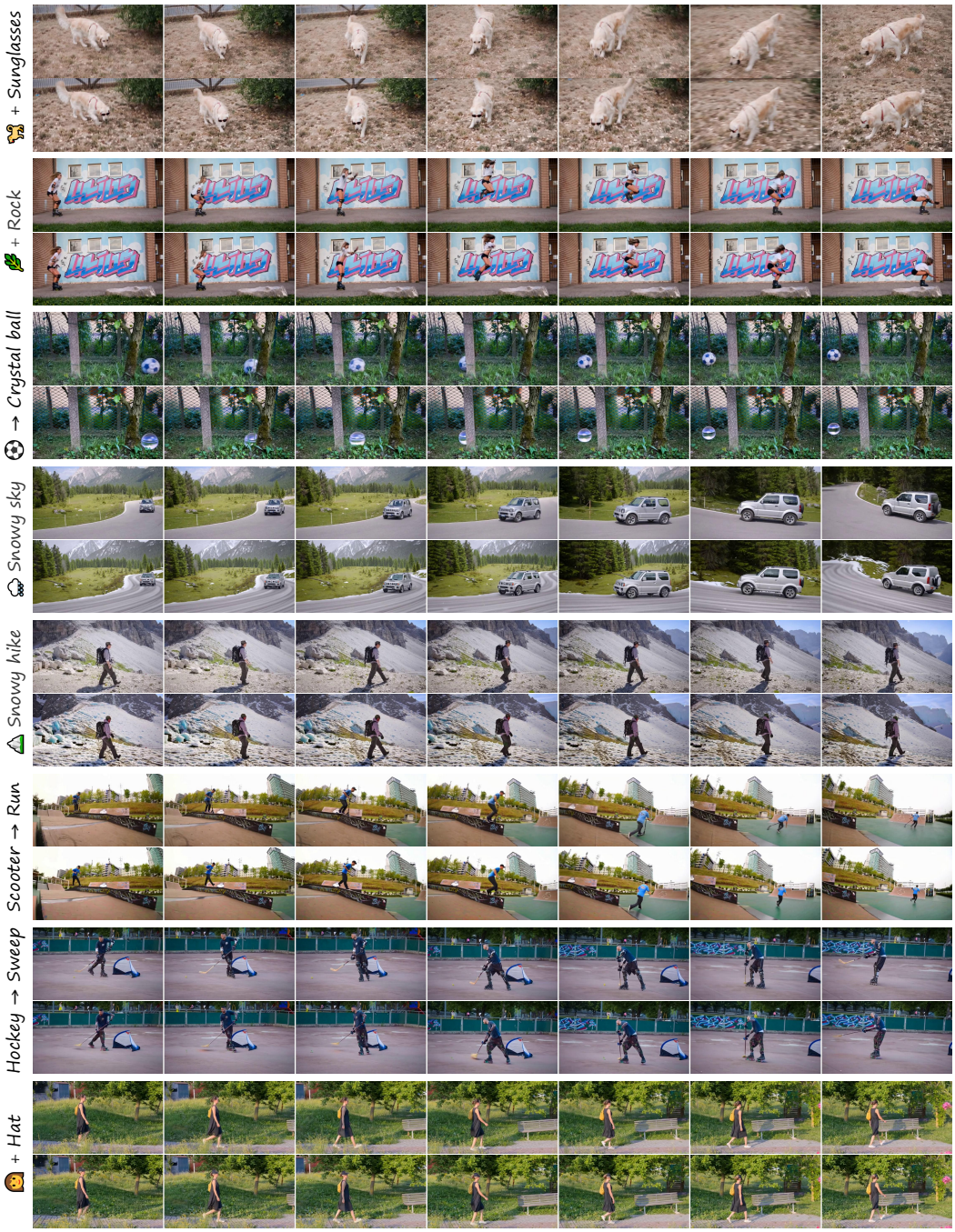


Figure I.15: **Additional qualitative results on video editing** on DAVIS dataset (Pont-Tuset et al., 2017). We set $\alpha = 0.8$, $\omega = 5$, $N = 25$ for the experiments.