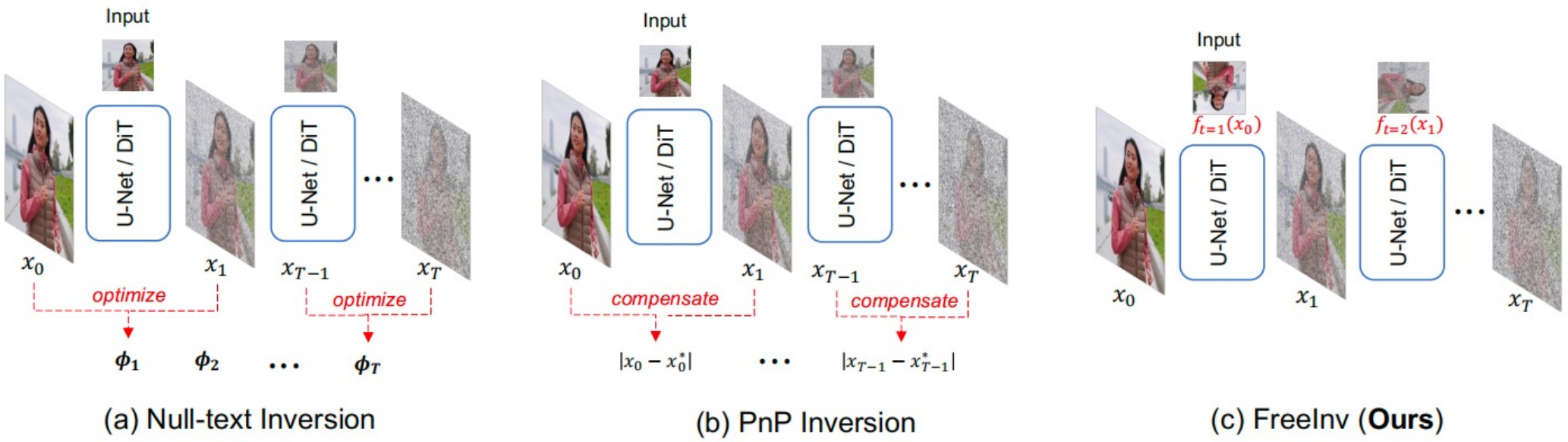
FreeInv: Free Lunch for Improving DDIM Inversion

Yuxiang Bao, Huijie Liu, Xun Gao, Huan Fu, Guoliang Kang

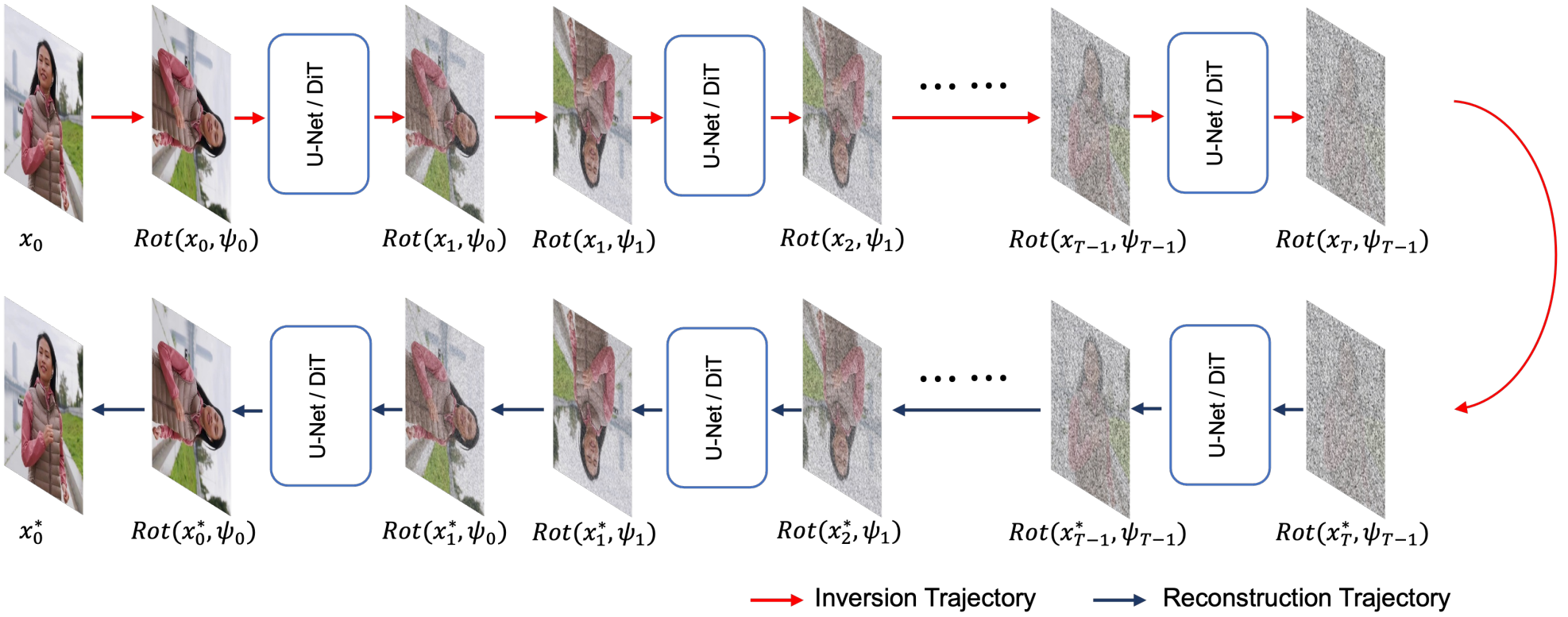
Code

Introduction



- Ideal DDIM inversion and reconstruction process is theoretically based on the local linear assumption, i.e., $\epsilon_\theta(x_t) \approx \epsilon_\theta(x_{t+1})$
- Learning-based methods aim to minimize the mismatch error through back-propagating the gradients to the null-text embedding
- Memory-based methods or store the errors generated in each time-step, and exploit them to make compensation.
- We propose a new method named **FreeInv** to deal with the trajectory deviation issue in a nearly free-lunch manner.

Approach



Detailed illustration of FreeInv. We employ rotation $Rot(\cdot,\cdot)$ as the transformation $f(\cdot)$ for example. During both the inversion and reconstruction phases, we rotate the latent representation with the same angle ψ_t at the t-th time-step, where ψ_t is randomly sampled. In practice, flipping, patch shuffling, jittering, etc., can serve as alternative.

Approach

How does FreeInv help reconstruction?

- **MBDI reduces mismatch error.** We propose to ensemble multiple trajectories of different images to enhance reconstruction fidelity by constructing a multi-branch DDIM inversion (MBDI) and reconstruction.

- **One-time MC sampling at each time-step** enables that randomly sampling a single branch at each timesteps brings performance comparable to that achieved with multiple MC samples or MBDI.

- **Image transformation as a branch**. To further improve efficiency, we replace the explicit multiple branches by applying transformations (e.g., rotation, flipping, patch-shuffling, etc.) to the image/latent, thereby generating multiple augmented versions.

Further mathematical details and derivations are provided in the paper.

Experiment

Methods		Reconstruction Accuracy				Inversion Computation Costs	
		PSNR ↑	LPIPS ($\times 10^{-2}$) ↓	MSE ($\times 10^{-3}$) ↓	SSIM ↑	Time (Seconds) ↓	Memory (MB) ↓
U-Net Based	DDIM Baseline [38]	25.04	9.14	4.43	0.77	4	**3031**
	NTI [25]	26.74	5.46	3.13	0.79	148	11945
	EDICT [45]	27.21	5.12	2.88	0.80	81	12325
	DI [15]	**28.19**	**4.76**	**2.29**	**0.81**	16	13595
	VI [50]	27.86	5.45	3.77	0.80	**3**	13853
	ReNoise [9]	26.61	6.52	3.19	0.79	21	6395
	BELM [46]	27.12	5.15	2.91	0.79	5	3641
	PI [17]	27.12	5.13	2.91	0.79	4	7197
	Ours	27.69	5.14	2.45	**0.81**	4	**3031**
DiT Based	FLUX [1]	14.92	38.60	46.19	0.54	**7**	**32430**
	FLUX+RF-Solver [47]	26.38	10.98	3.89	0.84	15	**32430**
	FLUX+**Ours**	**29.24**	**4.25**	**1.64**	**0.90**	**7**	**32430**

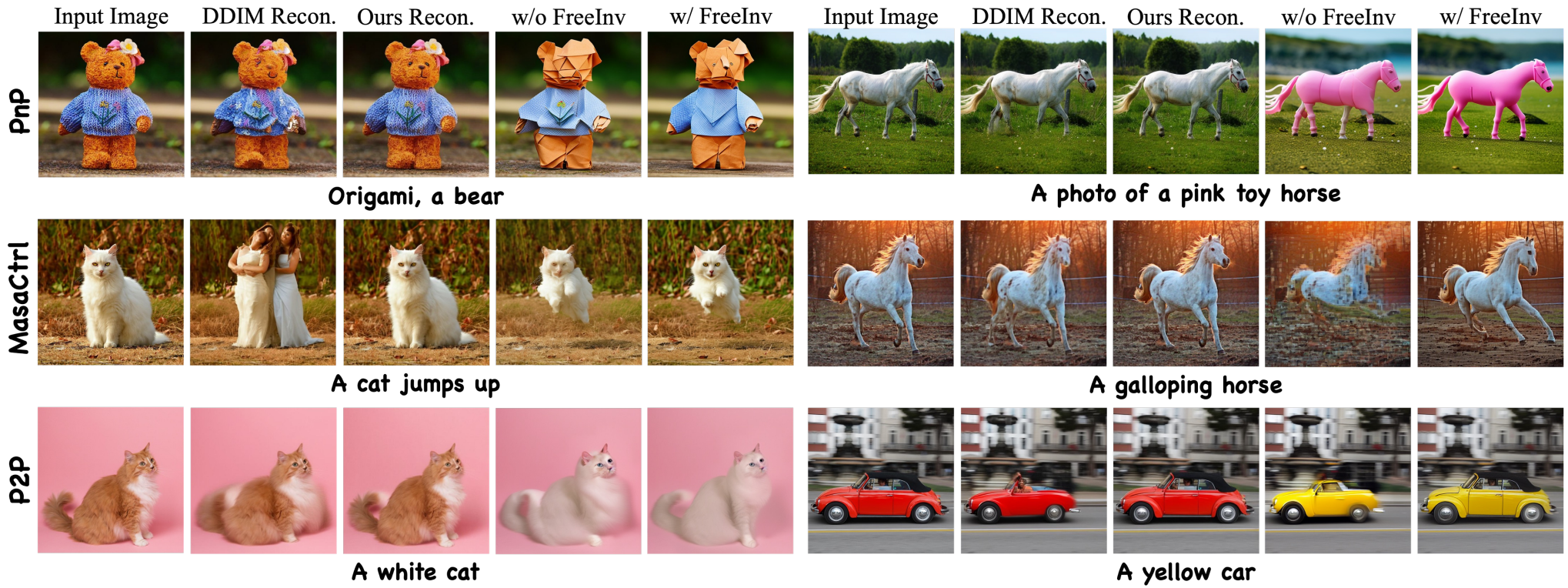
Quantitative comparison: reconstruction. We quantitatively evaluate reconstruction faithfulness, as well as computation costs of existing inversion methods, including U-Net based and DiT based methods, on the PIE benchmark. FreeInv achieves competitive results with superior high efficiency.

Experiment

Method		Structure Distance ($\times 10^{-3}$) ↓	Background Preservation				CLIP Similarity	
Inversion	Editing		PSNR ↑	LPIPS ($\times 10^{-2}$) ↓	MSE ($\times 10^{-3}$) ↓	SSIM ↑	Whole ↑	Edited ↑
DDIM [38]	P2P	69.88	17.84	21.02	22.07	0.71	25.18	22.33
NTI [25]	P2P	10.11	27.80	4.99	2.99	0.85	24.80	21.76
EDICT [45]	P2P	**3.84**	**29.79**	**3.70**	**2.04**	**0.87**	23.09	20.32
DI [15]	P2P	11.64	25.96	6.16	3.93	0.84	**25.60**	**22.61**
VI [50]	P2P	17.35	28.00	5.75	7.61	0.85	24.86	22.12
ReNoise [9]	P2P	23.25	25.11	8.97	5.14	0.82	23.81	21.16
BELM [46]	P2P	17.28	25.51	8.46	4.76	0.82	24.23	21.30
PI [17]	P2P	10.89	27.21	5.44	3.31	0.85	25.02	22.12
Ours	P2P	17.13	26.03	6.79	4.17	0.83	25.30	22.33

Quantitative comparison: editing. With P2P as baseline, we quantitatively compare existing inversion methods, with regard to background preservation and description alignment of edited images.

Experiment



Qualitative comparison. We integrate FreeInv into PnP, MasaCtrl, and P2P, respectively. We compare the reconstruction and editing results w/ or w/o FreeInv.

Experiment

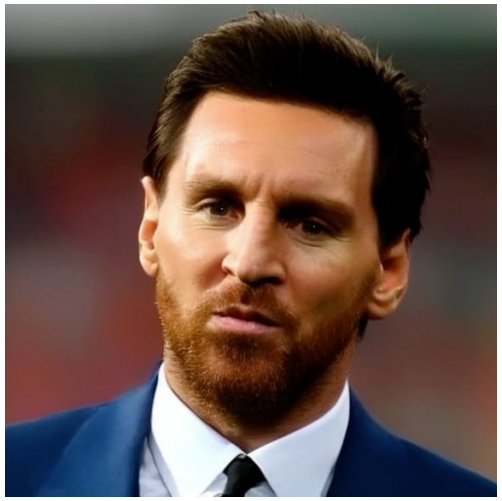
Video comparison. We compare TokenFlow, TokenFlow+STEM, and TokenFlow+Ours with respect to the reconstruction results (on the left side of the dash-line), the editing outcome (on the right side of the dash-line), as well as time and memory costs (below the reconstruction results).

Experiment

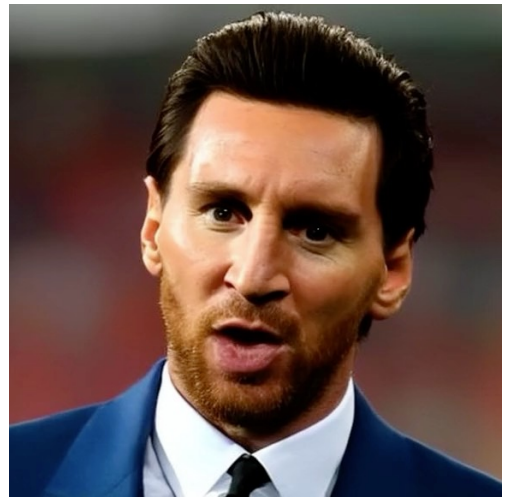
original video

DDIM recon.

FreeInv recon.



DDIM Editing



FreeInv Editing

Experiment

original video

DDIM recon.

FreeInv recon.

DDIM Editing

FreeInv Editing

Conclusion

In this paper, we find that an ensemble of trajectories for multiple images can effectively reduce the DDIM reconstruction error. Based on such a finding, we propose a method named FreeInv to perform an efficient ensemble. FreeInv enhances DDIM inversion in a free-lunch manner. In detail, we randomly transform the latent representation, and keep the transformation at each time-step the same between the inversion and the reconstruction. FreeInv is compatible with both U-Net and DiT architectures. Thanks to its efficiency, FreeInv is applicable not only to image reconstruction but also to video sequences. In both image and video reconstruction tasks, it achieves reconstruction fidelity comparable to or better than existing methods, while demonstrating significantly improved efficiency.