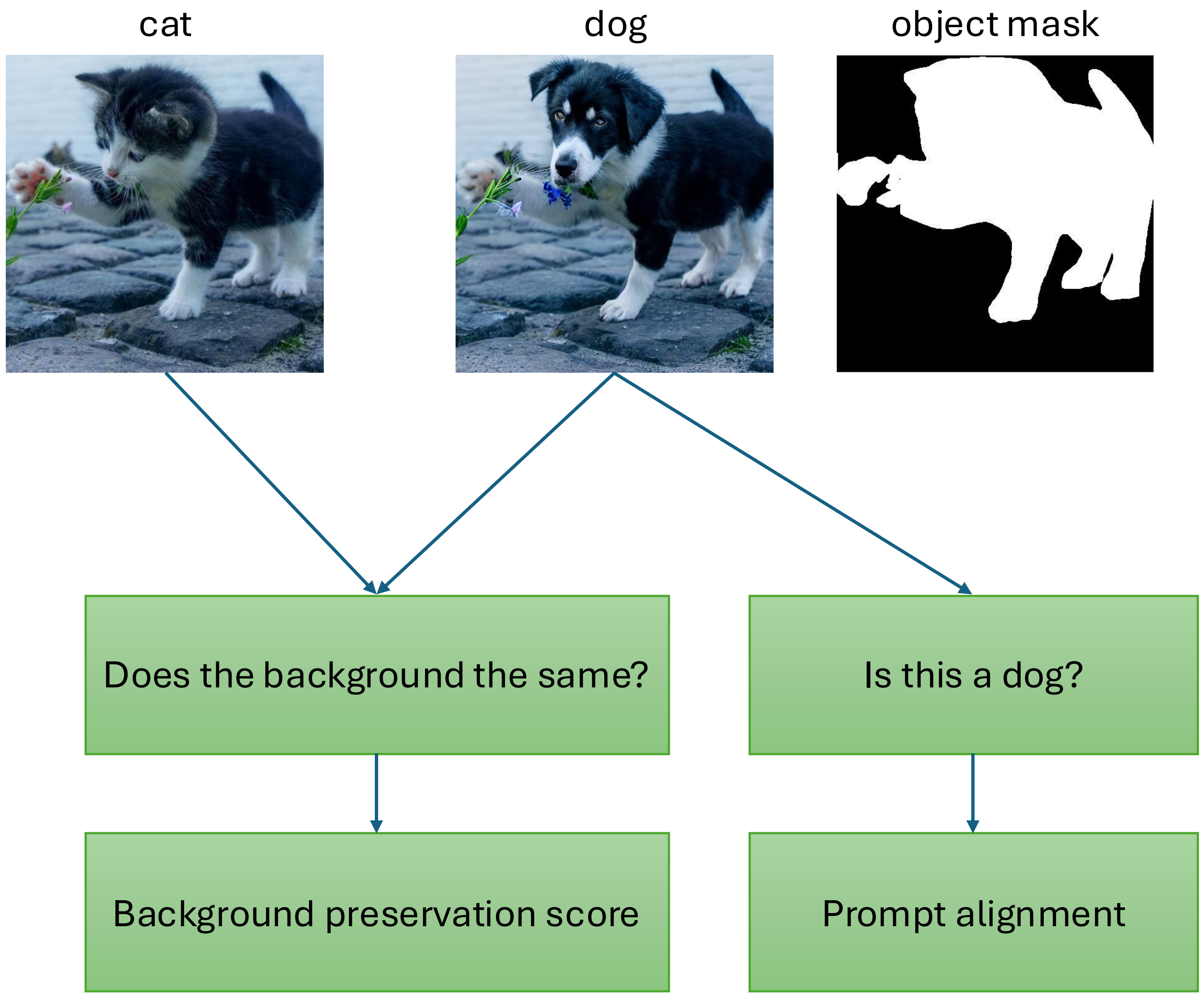
AutoEdit: Automatic Hyperparameter Tuning for Image Editing

Chau Pham, Quan Dao, Mahesh Bhosale, Yunjie Tian

Dimitris Metaxas, David Doermann

1. Text-based Image editing problem

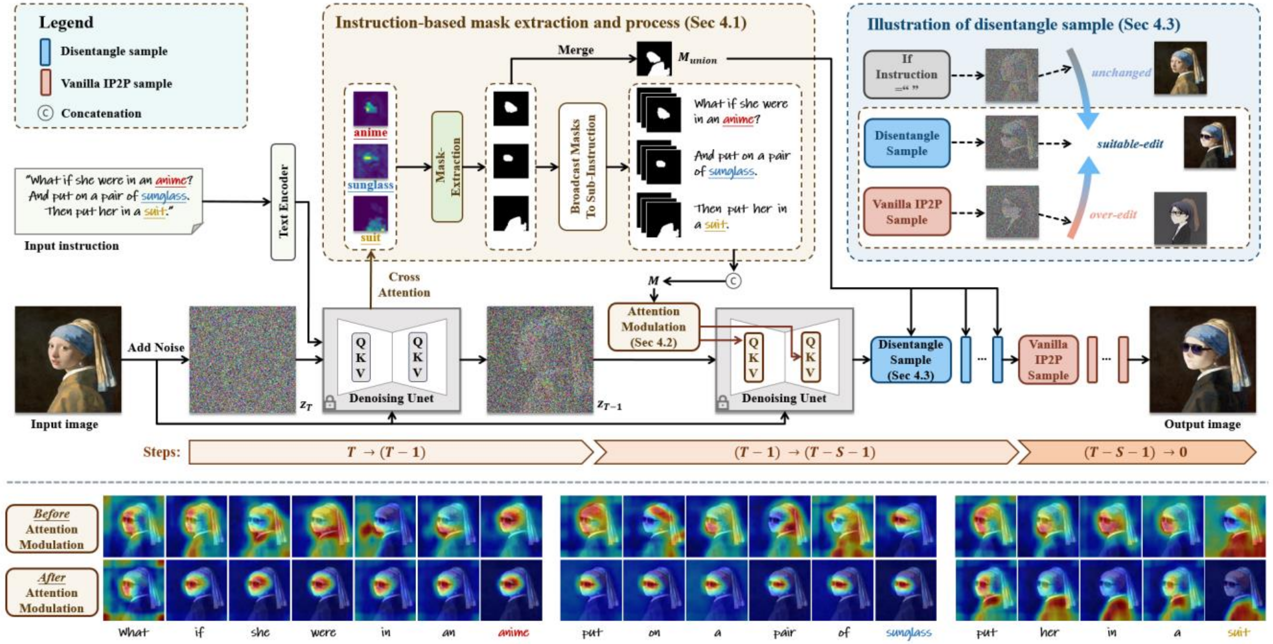


Two editing criteria:

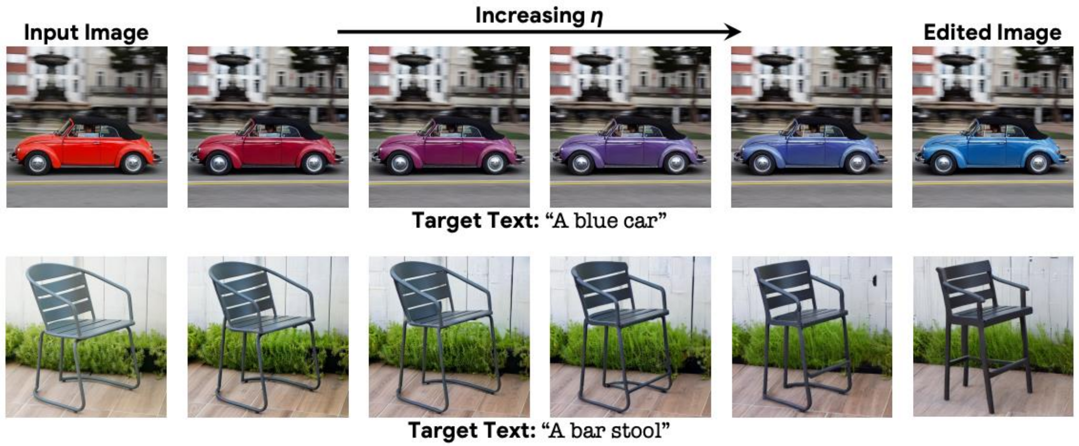
- Background preservation
- Prompt Alignment

2. Common image editing methods

NEURAL INFORMATION PROCESSING SYSTEMS



Attention control editing



Blending in latent space

Common approach:

- Inversion the image by applying an inversion method.
- Denoising: At each step of the denoising process, we need to choose the editing operation, decided by the hyperparameter.

Huang, Yuzhou, et al. "Smartedit: Exploring complex instruction-based image editing with multimodal large language models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Kawar, Bahjat, et al. "Imagic: Text-based real image editing with diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

3. Hyperparameter tuning for Image Editing

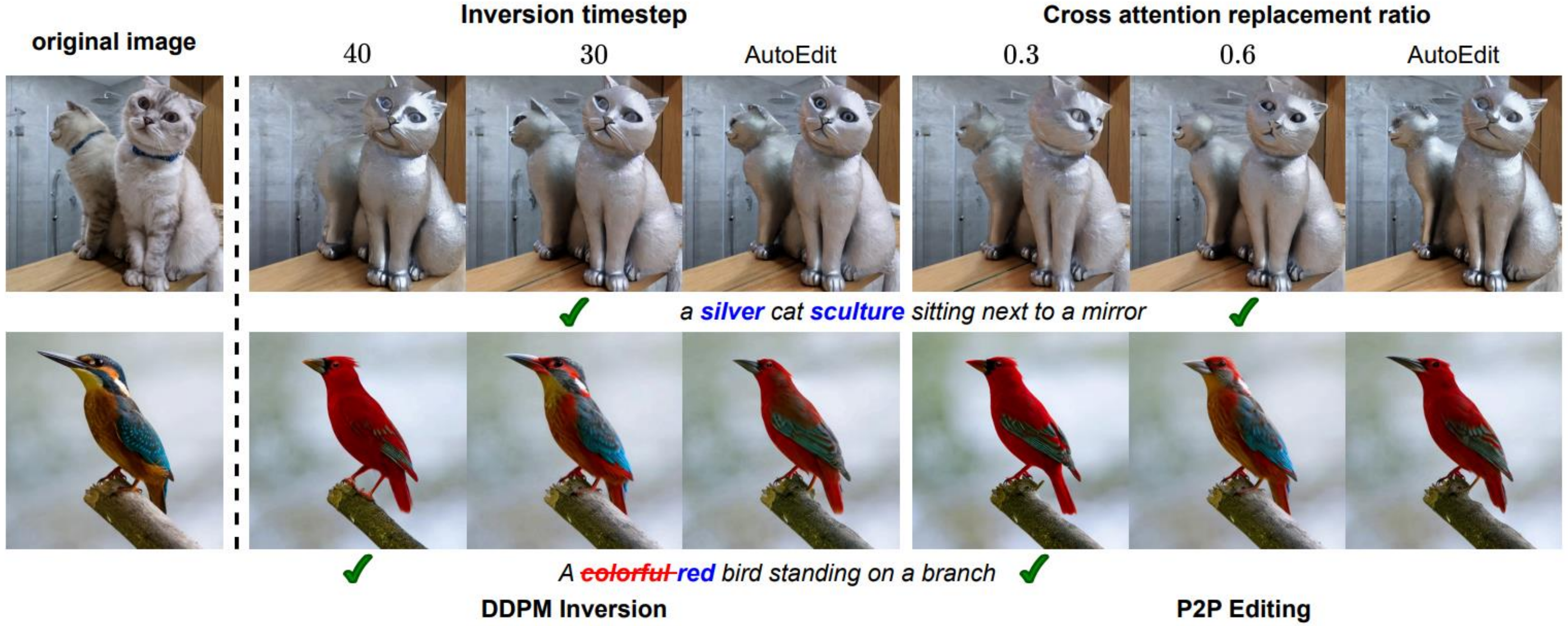
In Image Editing task, the programmers need to specify the hyperparameter:

- Inversion timestep
- Cross/Self attention ratio.
- Attention reweighting.
- Blending coefficient,…

The hyperparameters depend on the editing method. Each image has a different value of optimal hyperparameters.

Trial-and-error: If each hyperparameter can takes K values -> K times denoising to search the optimal value (O(TK) NFEs).

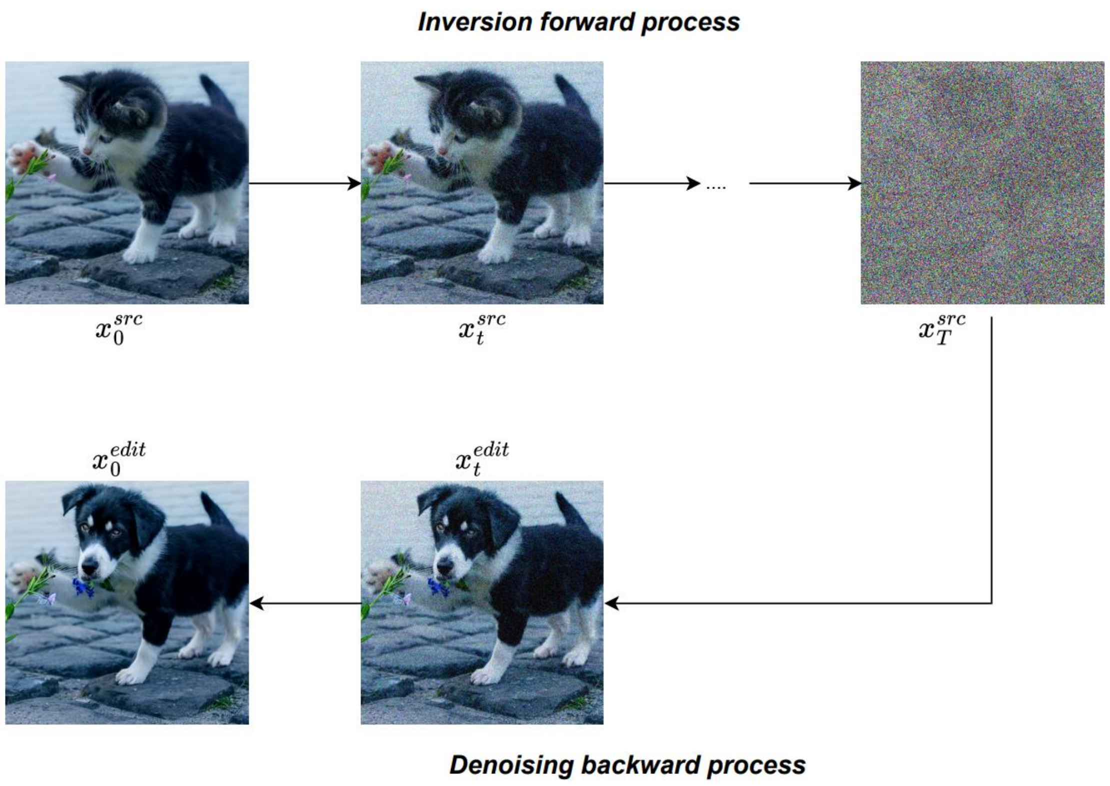
If there are N hyperparameters -> K^N times denoising -> (O(TK^N) NFEs)



4. Contribution

1. Discover the critical time-consuming of hyperparameter selection in trial-and-error methods.
2. Reformulating the hyperparameter searching in image editing as the RL problem -> applying PPO to train the RL.
3. The policy model can find near-optimal value of hyperparameters.

5. RL environment definition



RL is inserted in the denoising backward process:

- State: Noisy sample x_t. Initialize state at x_T
- Action: Parameterize the hyperparameter as the stepwise action H_t
- Reward: Consist of background preservation and prompt alignment.
- Termination: Finish after T steps.

6. AutoEdit Design

Reward function:

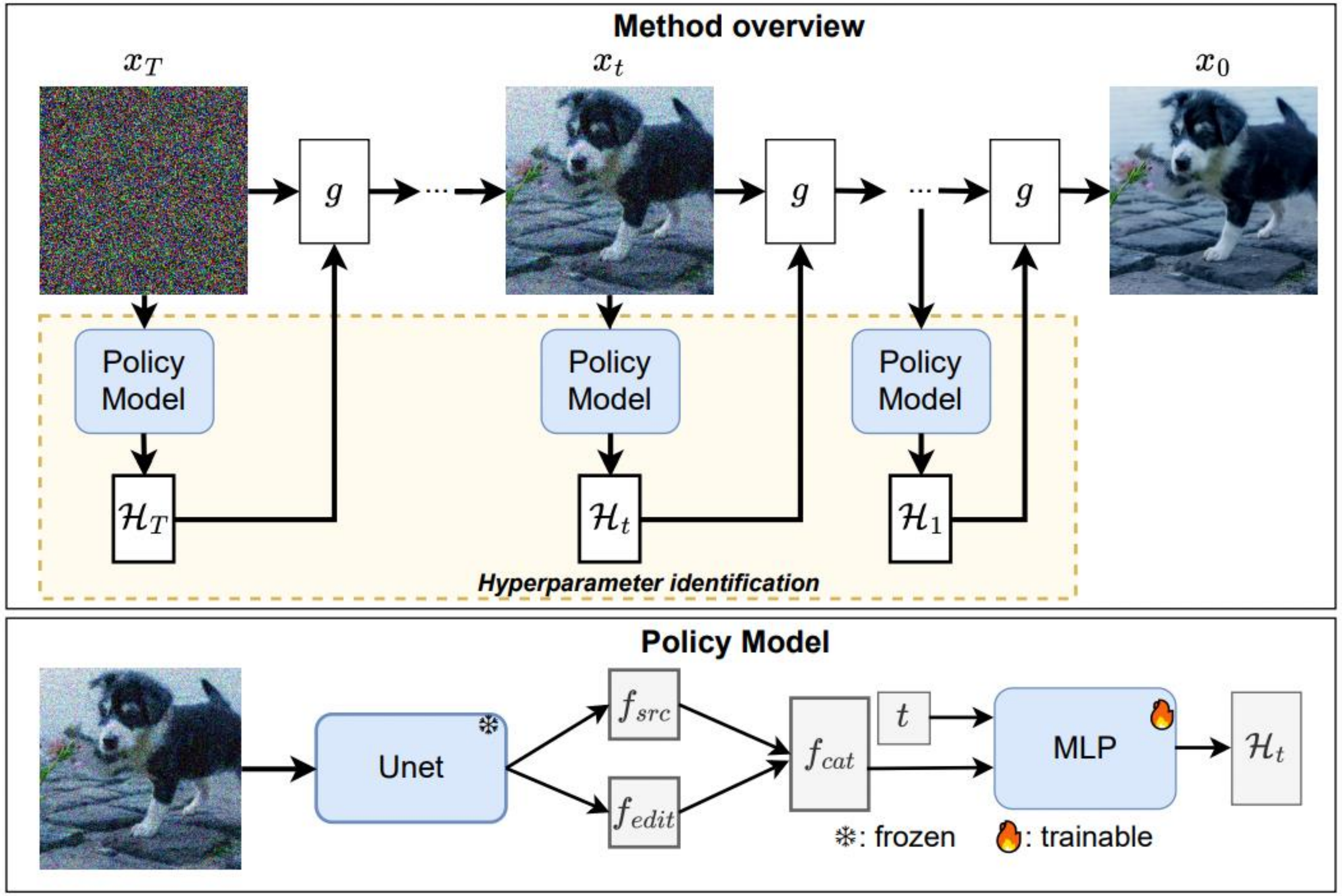
- Prompt alignment:
    - CLIP score of the edited region.
    - LLM judgement.
- Background preservation:
    - MSE score of unedited region.

Follow RL training for LLM, we conduct 2 stages:

- Policy initialization (SFT training)
- RL optimization.

Network design:

- Policy model: Use Unet encoder as feature extractor + several trainable layers for policy prediction.
- Value model: Similar with Policy model, but outputs a single scalar.



7. Experiments

Method	Base Model	Structure Distance ↓	Background Preservation PSNR ↑	SSIM ↑	MSE ↓	LPIPS ↓	CLIP Score Edited ↑	Whole ↑	LLM Score
DDIM-Inversion [39]	SD 1.4	38.10	21.36	76.67	103.95	146.60	**23.30**	**26.31**	0.96
+ AutoEdit		**18.74**	**24.65**	**81.28**	**52.94**	**95.10**	22.65	25.72	**1.12**
DDPM-Inversion [15]	SD 1.4	22.12	22.66	78.95	53.33	67.66	**23.02**	**26.22**	1.03
+ AutoEdit		**12.65**	**27.25**	**85.17**	**31.18**	**50.51**	22.52	25.83	**1.17**
PnP Inversion [16]	SD 1.5	11.65	27.22	84.76	35.86	60.67	22.10	25.02	1.10
+ AutoEdit		**11.06**	**27.85**	**85.04**	**33.77**	**60.12**	**23.00**	**25.79**	**1.19**
P2P [12]	SD 1.4	14.75	25.82	84.02	40.93	61.78	22.29	25.44	1.08
+ AutoEdit		**13.76**	**26.45**	**84.08**	**36.24**	**60.60**	**23.88**	**26.55**	**1.22**
MasaCtrl [5]	SD 1.4	28.38	22.17	79.67	86.97	79.67	21.16	23.96	0.92
+ AutoEdit		**21.33**	**23.48**	**80.06**	**46.28**	**71.35**	**21.75**	**24.86**	**0.99**
DDPM-Inversion [15]	SDXL	7.12	26.13	89.88	35.32	65.62	**23.0**	**27.11**	1.19
+AutoEdit		**6.46**	**27.86**	**90.50**	**20.44**	**53.51**	22.9	26.7	**1.27**
UltraEdit [47]	MM-DiT	10.82	26.5	84.7	46.7	75.8	22.4	25.6	1.20
+AutoEdit		**7.61**	**27.3**	**86.2**	**37.6**	**64.9**	**22.6**	**25.7**	**1.26**
InstructPix2Pix [4]	SD 1.5	35.37	20.8	76.4	226.8	157.3	22.1	24.5	0.65
+AutoEdit		**28.68**	**22.2**	**78.5**	**181.4**	**132.8**	**22.3**	**24.7**	**0.82**
Null-text [25]	SD 1.4	19.87	23.8	79.9	64.4	109.8	22.3	25.9	1.12
+AutoEdit		**10.91**	**25.7**	**82.4**	**45.4**	**82.3**	**22.6**	**26.3**	**1.21**

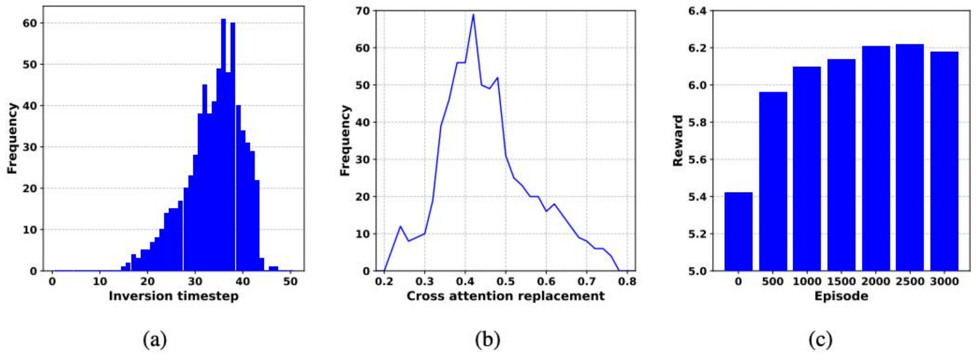
Method	PSNR	SSIM	CLIP Edit	CLIP Whole	LLM Score
Taming flow [42]	23.4	81.5	22.9	26.0	1.22
+AutoEdit	**25.7**	**85.2**	**23.4**	**26.1**	**1.30**
Fireflow [6]	23.1	82.2	22.4	25.2	1.20
+AutoEdit	**26.2**	**86.2**	**22.9**	**25.2**	**1.27**

NEURAL INFORMATION PROCESSING SYSTEMS

1. Generalize across editing methods.
2. Generalize across different Diffusion architecture.

7. Experiments

Qualitative Results



Test-time hyperparameter selection

7. Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

P1	P2	PSNR ↑	SSIM ↑	MSE ↓	LPIPS ↓	Edited ↑	Whole ↑	Reward
	✓	18.2	74.5	208.7	57.9	**23.2**	**26.3**	6.12
✓		22.1	77.4	52.7	69.7	20.7	23.4	5.42
✓	✓	**27.2**	**85.3**	**31.1**	**50.5**	22.5	25.8	**6.25**

Importance of Phase-1 training

α, β	PSNR ↑	SSIM ↑	MSE ↓	LPIPS ↓	Edited ↑	Whole ↑
$\alpha = 30, \beta = 10$	19.65	77.11	150.5	138.6	24.15	27.34
$\alpha = 30, \beta = 20$	23.59	82.15	66.84	82.30	23.44	26.95
$\alpha = 30, \beta = 30$	27.25	85.17	31.18	50.51	22.52	25.83
$\alpha = 30, \beta = 40$	28.53	86.03	24.72	42.80	21.36	24.36

Background preservation and prompt alignment tradeoff

Method	#Trials 1	2	3	AutoEdit	Optimal
DDIM-Inversion	5.81	6.03	6.11	6.09	6.17
DDPM-Inversion	6.11	6.21	6.23	6.25	6.32
P2P	6.17	6.31	6.37	6.38	6.45
MasaCtrl	5.47	5.59	5.65	5.65	5.75

Convergence of AutoEdit

Method	PSNR	SSIM	MSE	LPIPS	Edited	Whole	LLM
DDPM Inv	26.1	89.8	35.3	65.6	23.0	27.1	1.19
+ AutoEdit	27.8	90.5	20.4	53.5	22.9	26.7	1.27
+ AutoEdit + LLM	**29.1**	**91.8**	**19.1**	**49.1**	22.7	26.6	**1.31**

LLM Score as reward function

Thank you