Boundary Guided Learning-Free Semantic Control with Diffusion Models

Ye Zhu[1,2], Yu Wu[3], Zhiwei Deng[4], Olga Russakovsky[2], Yan Yan[1]

[1]Illinois Institute of Technology, USA

[2]Princeton University, USA

[3]Wuhan University, China

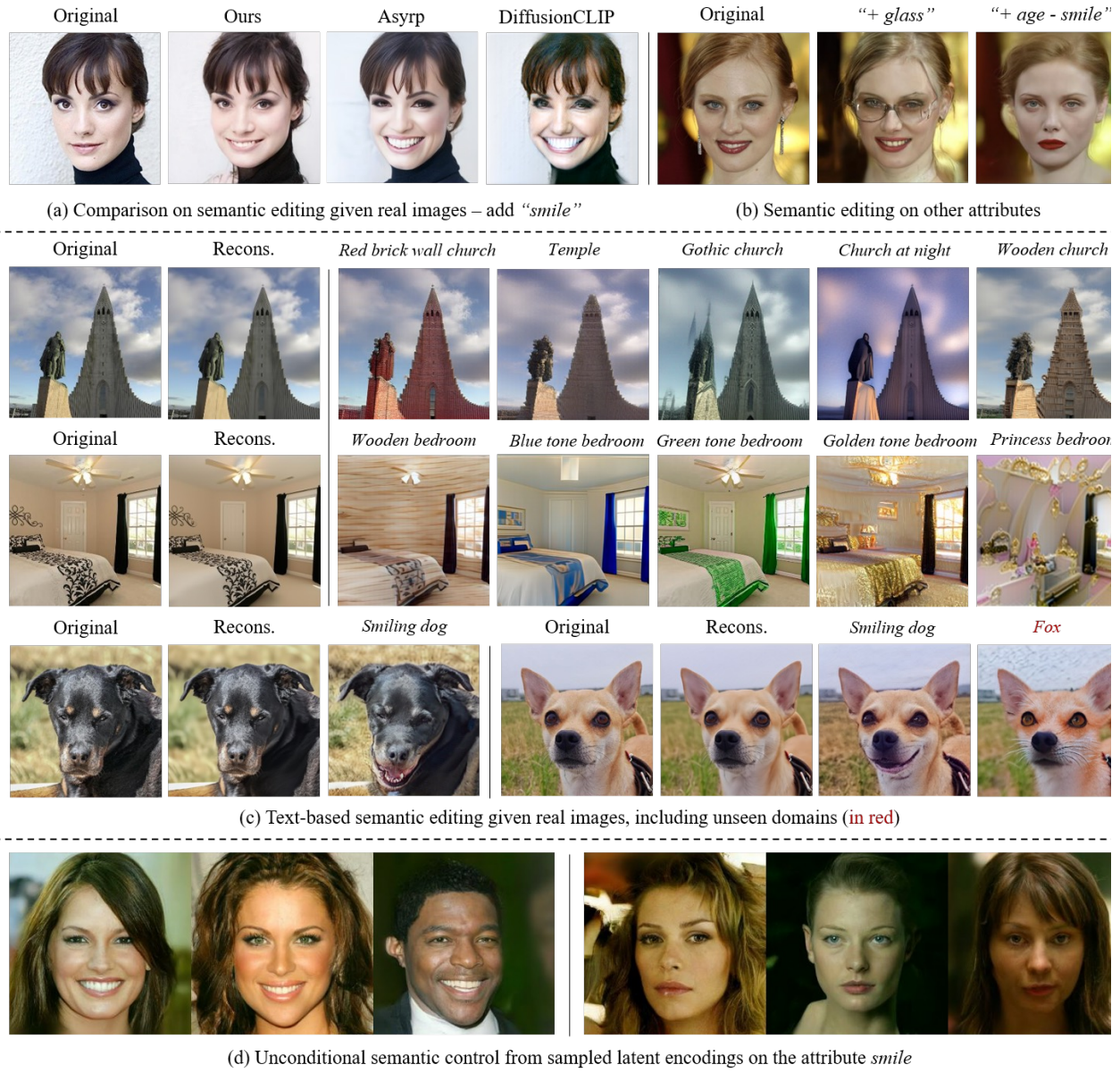
[4]Google Research, USA

Paper

Code

BoundaryDiffusion

TL;DR: We introduce one of the first **learning-free** methods for effective and efficient image semantic control and editing via pre-trained and frozen **unconditional** denoising diffusion base models.



(a) Comparison on semantic editing given real images – add *"smile"*

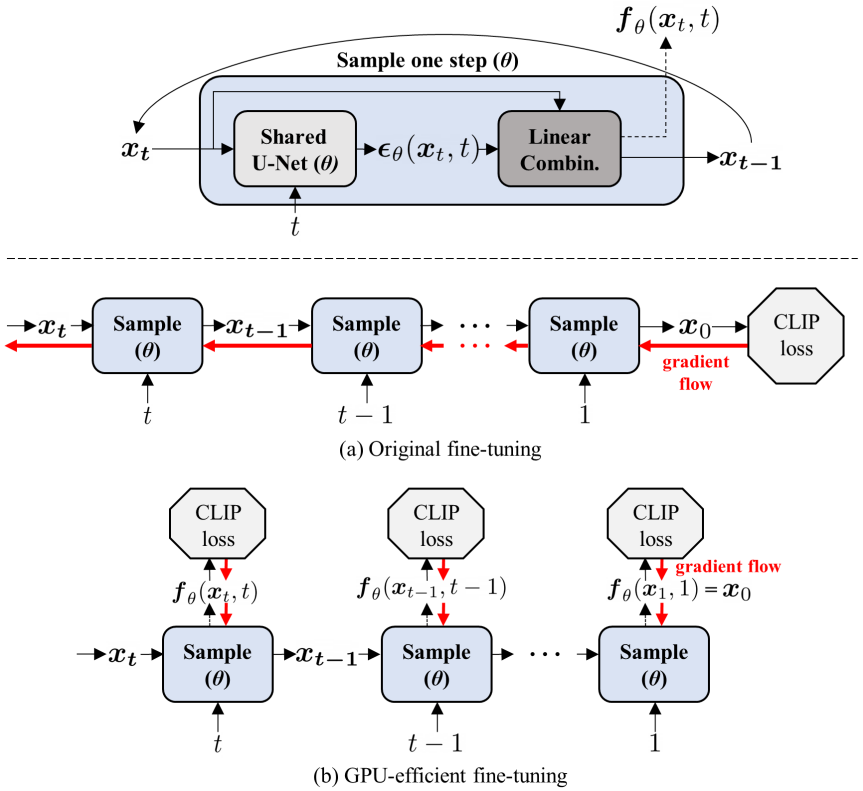
(b) Semantic editing on other attributes

(c) Text-based semantic editing given real images, including unseen domains (in red)

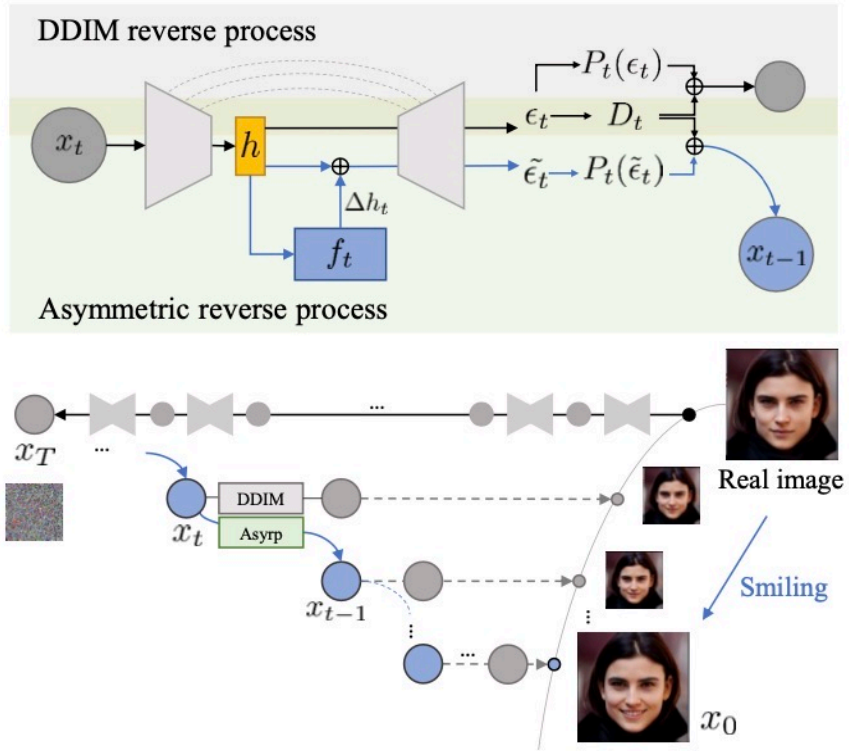
(d) Unconditional semantic control from sampled latent encodings on the attribute *smile*

Problem Overview and Current Paradigms

Unconditional diffusion models have achieved impressive performance in image synthesis, but are usually considered to be less semantic-aware in the generic noisy latent spaces.



DiffusionCLIP [CVPR 22']: fine-tuning



Asyrp [ICLR 23']: auxiliary networks

Distance Effect – distorted images by direct latent editing

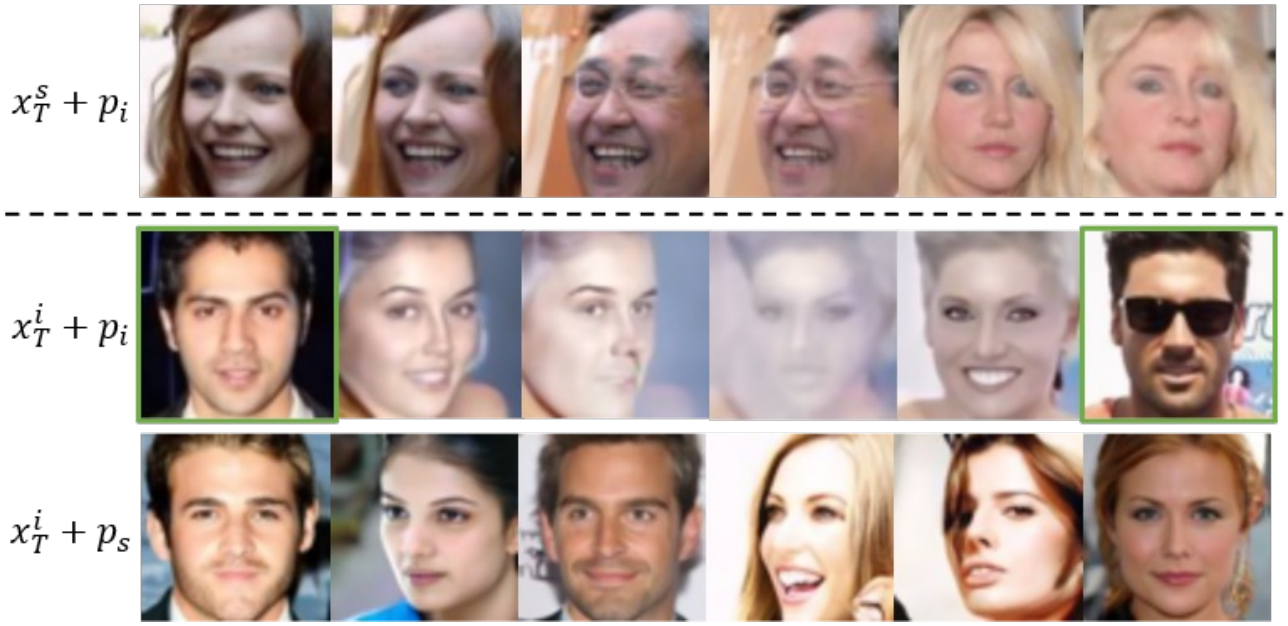


Figure of the Distance Effect: following the same denoising process, inverted latent encodings lead to distorted images.

Take-away: Distance effect indicates that the inversion trajectory from image to latent space via DDIMs [*Song et al., ICLR 21*] is asymmetric to denoising trajectory, which contradicts to existing works [*Kwon et al., ICLR 23*].

Gaussian Spaces – geometric and probabilistic properties

Take-away: By using the standard Gaussian space and its established properties as the reference, we confirm our asymmetric assumption, and propose to further dissect the denoising trajectory to trace the critical step(s) that are suitable for controlling semantics.

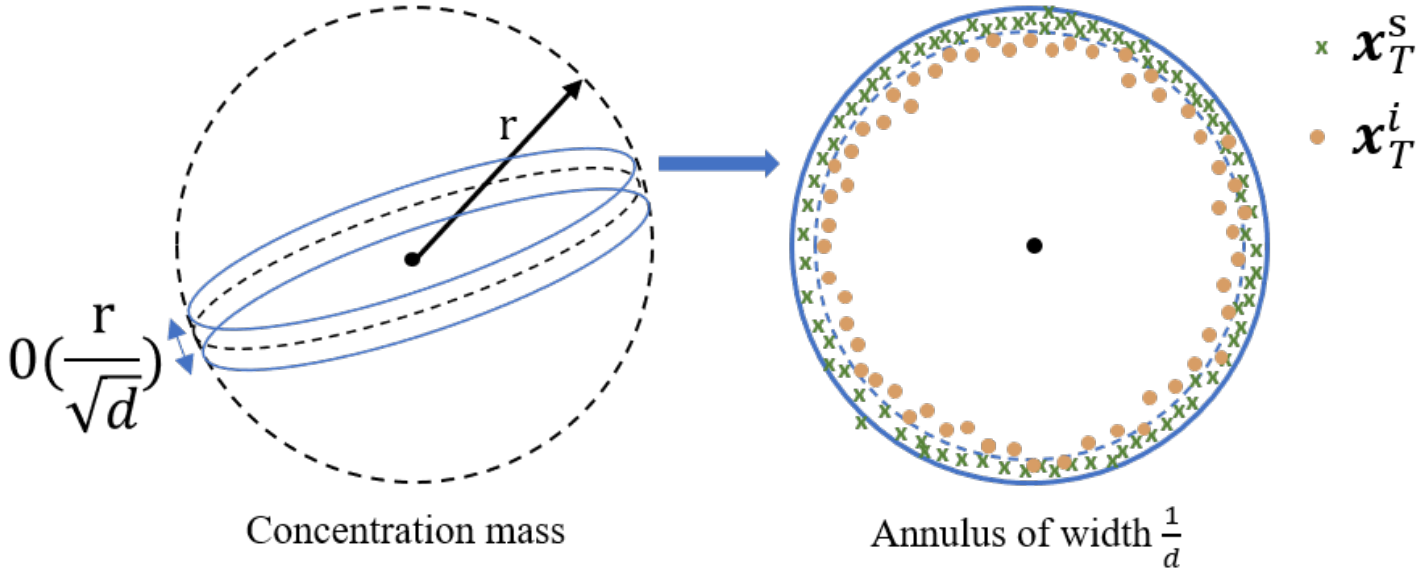


Figure of the Gaussian Sphere: geometric illustration of the concentration mass in high-dimensional Gaussian.

BoundaryDiffusion Design

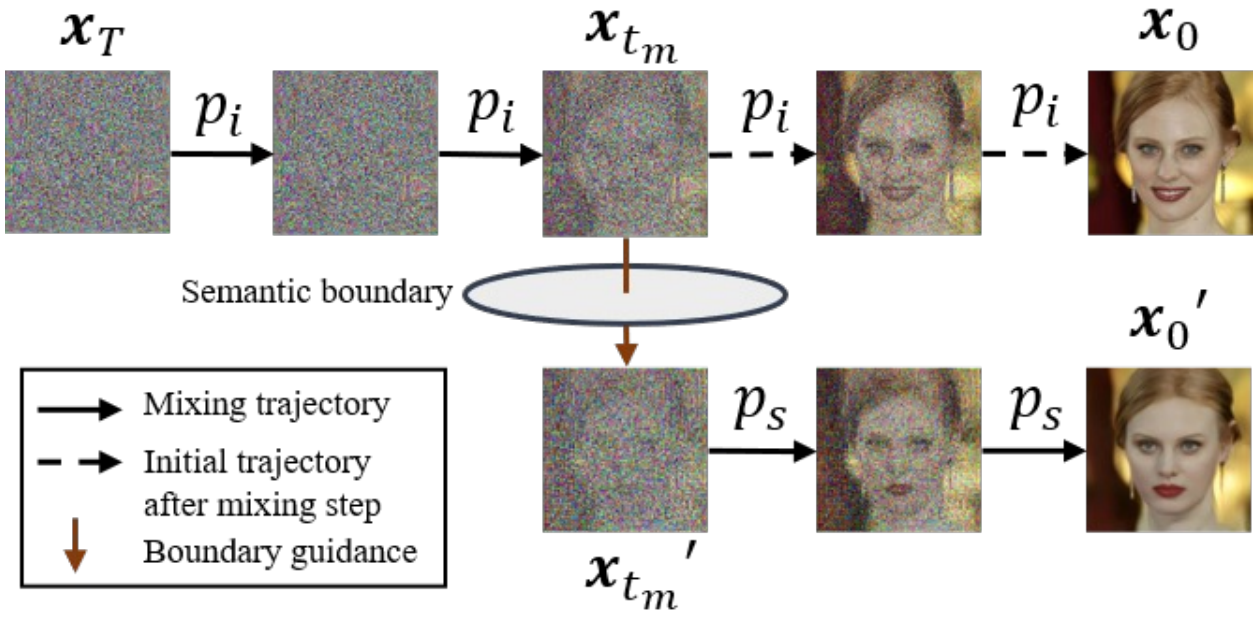


Figure of the *BoundaryDiffusion* for semantic control in one-step editing.

Step 1- Inversion: Get raw images (~100) with known semantics and invert them into the mixing step (see details about how to find the mixing step in our paper)

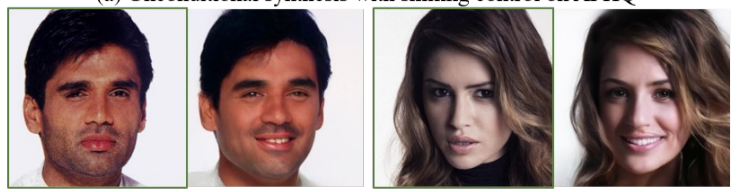
Step 2 - Boundary: Find the boundary hyperplanes with SVMs (no learning, just fit a SVM to localize pre-formed semantic boundaries) in ~ 1 second.

Step 3- Editing: Impose boundary-guided editing in the mixing step.

Step 4- Denoising: Follow the stochastic denoising to get edited images.

Experiments – versatile and non-cherry-picky

(a) Unconditional synthesis with smiling control on AFHQ



(b) Real image conditioned editing with smiling control on CelebA-HQ

(c) Text-based image editing with "*red brick wall church*" on LSUN-Church

Table 3: Evaluation results on CelebA-HQ-256 for real image conditioned semantic editing. FID scores are reported on the test set with 500 raw images, averaged on *"add or remove smile"* editing. The user study follows similar evaluation questions in [35].

Methods	S_{dir} ↑	SC ↑	ID ↑	FID ↓	User Quality ↑	User Attribute ↑
StyleCLIP [43]	0.13	86.8%	0.35	-	-	-
StyleGAN-NADA [15]	0.16	89.4%	0.42	-	-	-
DiffusionCLIP [31]	0.17	**93.7%**	0.70	86.23	3.2%	8.2%
Asyrp [35]	**0.19**	87.9%	-	68.38	41.3%	44.9%
Ours BoundaryDiffusion	0.17	90.4%	**0.73**	**63.14**	**55.5%**	**46.9%**

Thank you !

Poster Info:
Great Hall & Hall B1+B2 #214
Wed 13 Dec 8:45 a.m. PST — 10:45 a.m. PST
New Orleans, USA

Paper

Code