ArcDAE

Asymmetric Rectified Contrastive Diffusion Autoencoder for Unified Representation Learning

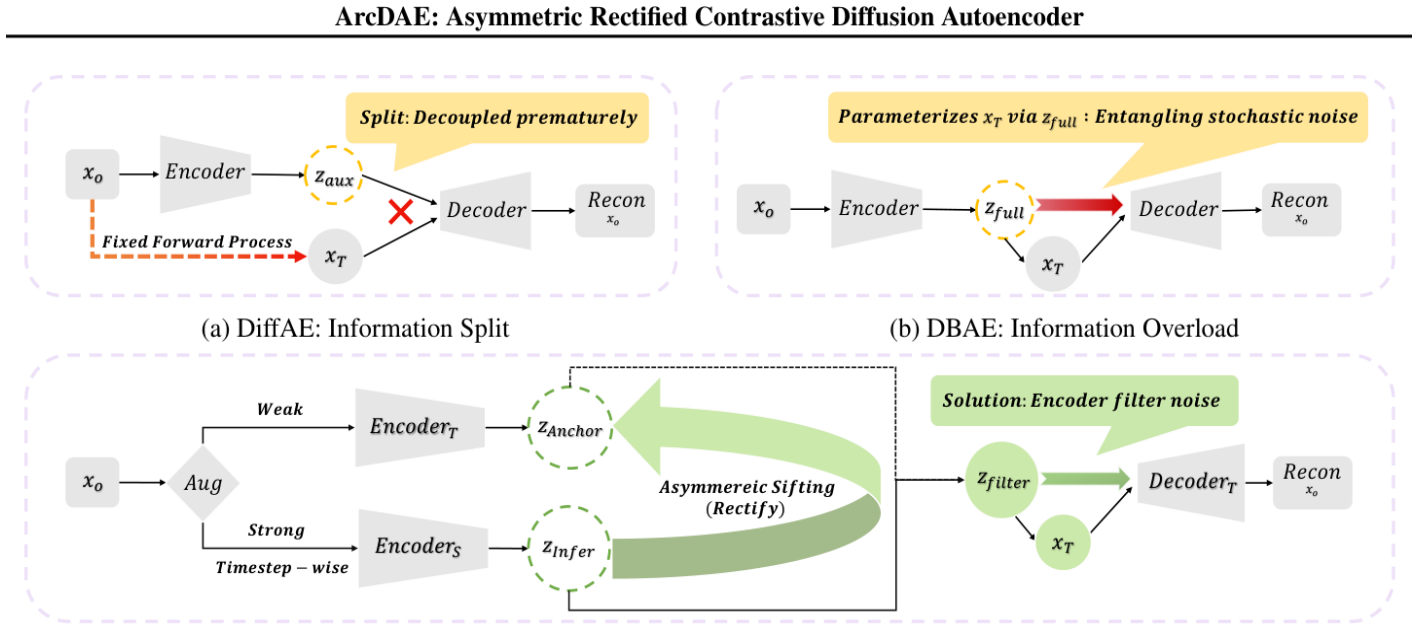
Ge Gao · Di Xiong · Zeke Xie · Jian Yang · Shuo Chen*

Nanjing University · HKUST (Guangzhou) · Nanjing University of Science and Technology

ICML 2026 · Seoul, South Korea

Limitations of Existing Works

Two structural extremes: information split and information overload.



Paper Fig. 1: conceptual comparison

Existing designs oscillate between two failures

Information split

Auxiliary semantic code is decoupled from the generative driver, limiting representational completeness.

Information overload

Bridge bottlenecks reconstruct noisy endpoints and entangle high-frequency stochastic variation.

understanding invariance → reconstruction fidelity

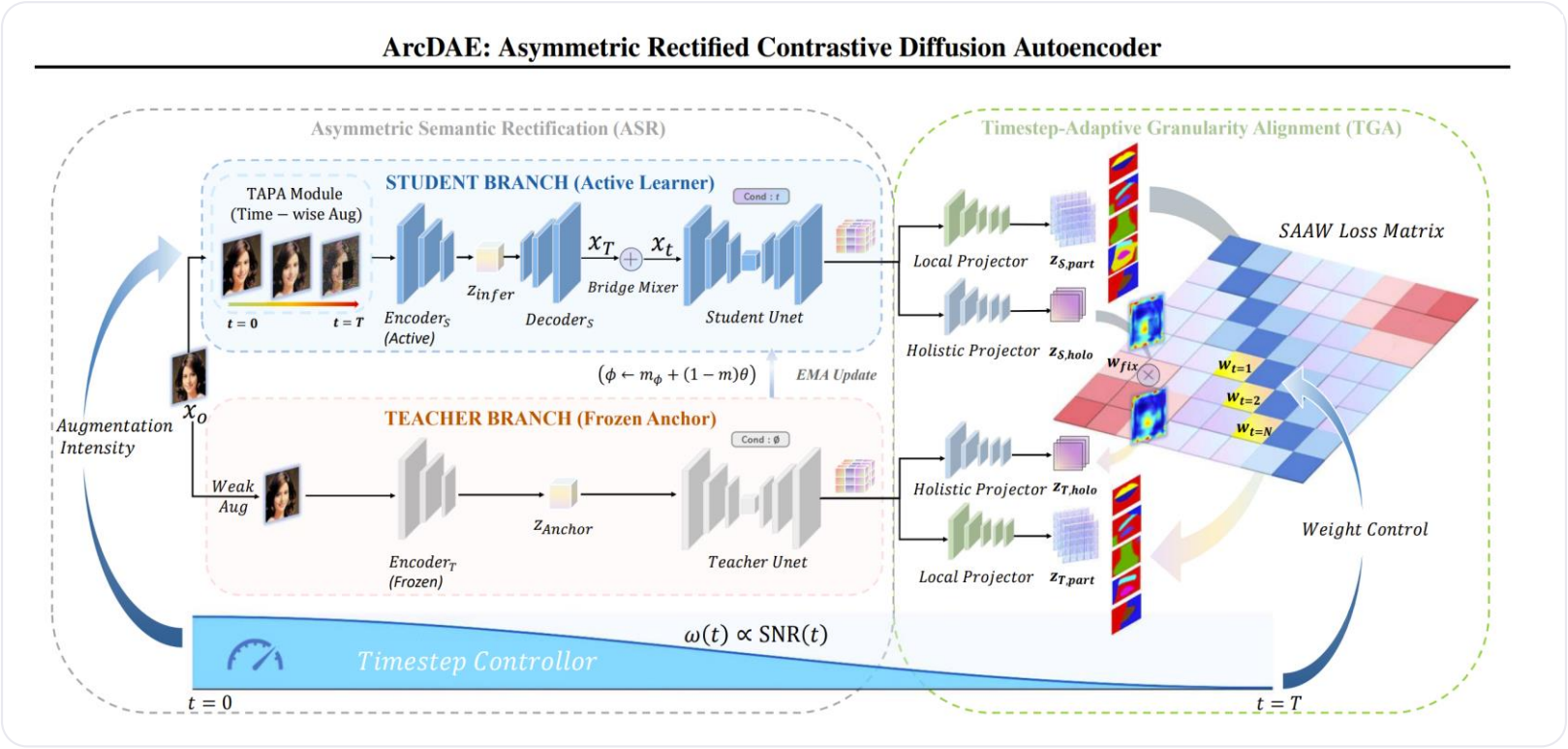
ArcDAE targets the high-fidelity semantic region

Core hypothesis

A robust encoder should act as an information sifter, not a passive reconstruction bridge.

Our Proposed Method: ArcDAE

Rebuilding the diffusion bridge as a dynamic information sifter.



Paper Fig. 2: asymmetric teacher–student framework

Three design ingredients

- ASR: a clean teacher anchor rectifies the noisy student representation.
- TAPA: augmentation strength increases with diffusion timestep.
- TGA: global and local alignment are gated by signal-to-noise ratio.

Clean semantic anchor → noisy student inference → timestep-aware contrastive rectification

Objective: Rectify Semantics, Gate Granularity

The loss balances diffusion reconstruction with contrastive semantic alignment.

ASR Smoothness Effect

High-noise timesteps strengthen contraction against unstable nuisance variation.

$$\begin{cases} \mathbf{z}^{\mathcal{T}} = \mathrm{sg}\left[E_\phi(\mathbf{v}^{\mathcal{T}})\right], & \mathbf{v}^{\mathcal{T}} \sim p_{\mathrm{aug}}(\cdot \mid \mathbf{x}_0, 0), \\ \mathbf{x}_t^{\mathcal{S}} = \alpha_t \mathbf{v}^{\mathcal{S}} + \sigma_t \boldsymbol{\epsilon}, & \mathbf{v}^{\mathcal{S}} \sim p_{\mathrm{aug}}(\cdot \mid \mathbf{x}_0, \gamma(t)), \end{cases}$$

$$\mathcal{L}_{\mathrm{ASR}} = \underbrace{\left\| E_\theta(\mathbf{m}_t) - \mathbf{z}^{\mathcal{T}} \right\|_2^2}_{\textit{Alignment error}} + \underbrace{\sigma_t^2 \,\mathrm{Tr}\left(\mathbf{J}_{E_\theta}(\mathbf{m}_t)^\top \mathbf{J}_{E_\theta}(\mathbf{m}_t)\right)}_{\textit{Jacobian regularization}} + \mathcal{O}\left(\sigma_t^2 \left\| E_\theta(\mathbf{m}_t) - \mathbf{z}^{\mathcal{T}} \right\|_2 \left\| \mathbf{H}_{E_\theta}(\mathbf{m}_t) \right\|_{\mathrm{op}} + \sigma_t^4\right).$$

Holistic & Local Representations

$$\mathbf{z}_{\mathrm{part}} = \mathcal{P}_{\mathrm{part}}\left(\int_\Omega \mathbf{h}(\mathbf{u})\, \mathrm{d}\rho_{\mathbf{q}}(\mathbf{u})\right). \qquad Z_\rho(\mathbf{q}) := \int_\Omega \exp\left(\frac{\langle \mathbf{q}, \phi_K(\mathbf{h}(\mathbf{u}))\rangle}{\sqrt{d_k}}\right) \mathrm{d}\mu(\mathbf{u}),$$

$$\mathbf{z}_{\mathrm{holo}} = \mathcal{P}_{\mathrm{holo}}\left(\int_\Omega \mathbf{h}(\mathbf{u})\, \mathrm{d}\mu(\mathbf{u})\right). \qquad \frac{\mathrm{d}\rho_{\mathbf{q}}}{\mathrm{d}\mu}(\mathbf{u}) = \frac{1}{Z_\rho(\mathbf{q})} \exp\left(\frac{\langle \mathbf{q}, \phi_K(\mathbf{h}(\mathbf{u}))\rangle}{\sqrt{d_k}}\right).$$

TGA Module : Timestep-Aware Adaptive Weighting

t → 0: local structure preserved · t → T: local noise alignment attenuated

$$\omega(t) \equiv \omega_{\tau_{\mathrm{snr}}}(t) := \frac{\mathrm{SNR}(t)^{1/\tau_{\mathrm{snr}}}}{1 + \mathrm{SNR}(t)^{1/\tau_{\mathrm{snr}}}}, \qquad \tau_{\mathrm{snr}} > 0.$$

Total training objective

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{diff}}(\theta) + \lambda_{\mathrm{align}}\left[\ell_{\mathrm{NCE}}(\mathbf{z}^{\mathcal{S}}_{\mathrm{holo}}, \mathbf{z}^{\mathcal{T}}_{\mathrm{holo}}) + \omega(t) \cdot \ell_{\mathrm{NCE}}(\mathbf{z}^{\mathcal{S}}_{\mathrm{part}}, \mathbf{z}^{\mathcal{T}}_{\mathrm{part}})\right].$$

Interpretation

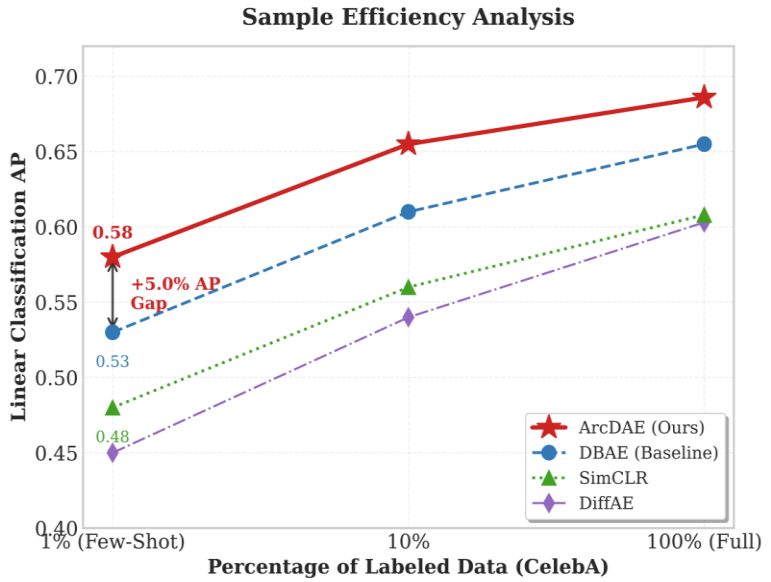
The diffusion loss keeps reconstructive ability; the gated contrastive term prevents the latent code from becoming a noise map.

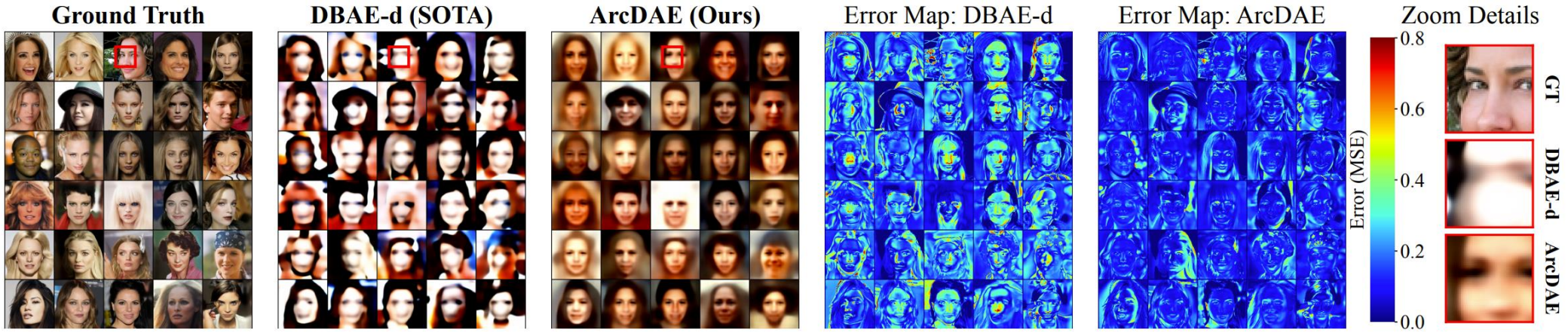
data-like high SNR → preserve local structure → sift stochastic noise → noise-like low SNR

Experimental Setup & Results & Analysis I

Evaluation covers semantic density, reconstruction, generation, and ablations.

Datasets	CelebA · FFHQ · CelebA-HQ; additional non-face benchmarks
Backbone	ADM U-Net encoder / score network; latent dimension dz = 512
Baselines	SimCLR, β-TCVAE, IB-GAN, DiffAE, PDAE, DiTi, DBAE
Metrics	AP, Pearson's r, MSE, SSIM, LPIPS, FID, Precision / Recall, TAD





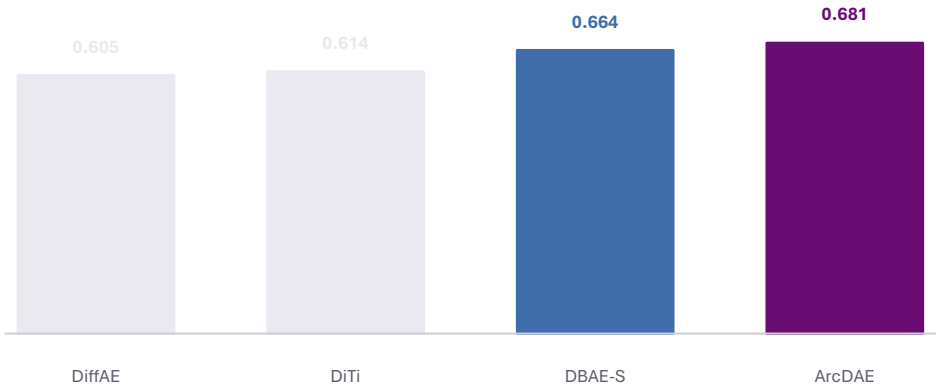
Early convergence: fewer artifacts and lower error maps

Main question: Does sifting improve semantics without sacrificing generative fidelity?

Results & Analysis II

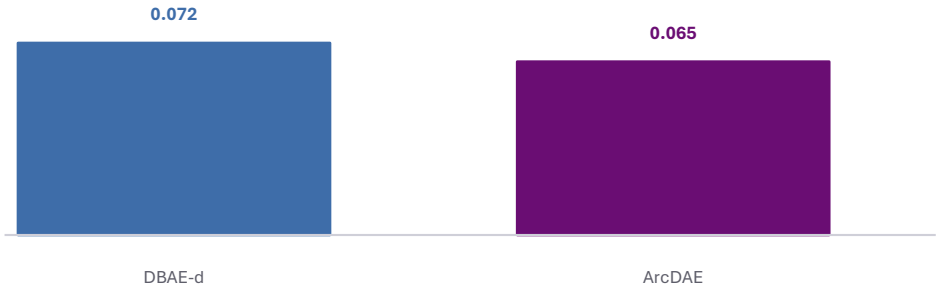
ArcDAE improves semantic density and perceptual reconstruction.

Linear-probe AP on FFHQ



Method	Gen	CelebA AP (↑)	CelebA Pearson's r (↑)	CelebA MSE (↓)	FFHQ AP (↑)	FFHQ Pearson's r (↑)	FFHQ MSE (↓)
SimCLR (Chen et al., 2020)	✗	0.597	0.474	0.603	0.608	0.481	0.638
β-TCVAE (Chen et al., 2018)	✓	0.450	0.378	0.573	0.432	0.335	0.608
IB-GAN (Jeon et al., 2021)	✓	0.442	0.307	0.597	0.428	0.260	0.644
DiffAE (Preechakul et al., 2022)	✓	0.603	0.598	0.421	0.605	0.606	0.410
PDAE (Zhang et al., 2022)	✓	0.602	0.596	0.410	0.597	0.603	0.416
DiTi (Yue et al., 2024)	✓	0.623	0.617	0.392	0.614	0.622	0.384
DBAE (Det.) (Kim et al., 2025)	✓	0.650	0.635	0.413	0.656	0.638	0.404
DBAE (Stoch.) (Kim et al., 2025)	✓	0.655	0.643	0.369	0.664	0.675	0.332
ArcDAE (Ours)	✓	**0.676**	**0.656**	**0.345**	**0.681**	**0.718**	**0.319**

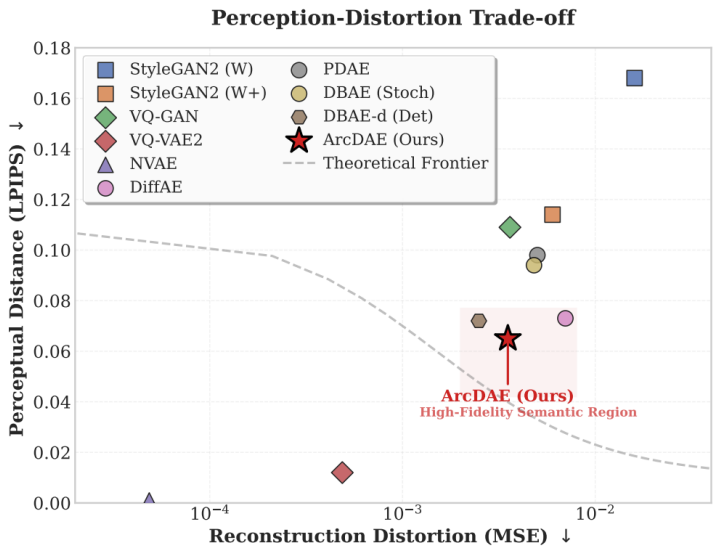
LPIPS reconstruction on CelebA-HQ ↓



Method	Tractability	Latent Dim (↓)	SSIM (↑)	LPIPS (↓)	MSE (↓)
StyleGAN2 ($\mathcal{W}$) (Karras et al., 2020)	✗	512	0.677	0.168	0.016
StyleGAN2 ($\mathcal{W}+$) (Abdal et al., 2019)	✗	7,168	0.827	0.114	0.006
VQ-GAN (Esser et al., 2021)	✓	65,536	0.782	0.109	3.61×10^{-3}
VQ-VAE2 (Razavi et al., 2019)	✓	327,680	0.947	0.012	4.87×10^{-4}
NVAE (Vahdat & Kautz, 2020)	✓	6,005,760	0.984	0.001	4.85×10^{-5}
DDIM (Inferred $\mathbf{x}_T$) (Song et al., 2021a)	✗	49,152	0.917	0.063	0.002
DiffAE (Inferred $\mathbf{x}_T$) (Preechakul et al., 2022)	✗	49,664	0.991	0.011	6.07×10^{-5}
PDAE (Inferred $\mathbf{x}_T$) (Zhang et al., 2022)	✗	49,664	0.994	0.007	3.84×10^{-5}
DiffAE (Random $\mathbf{x}_T$) (Preechakul et al., 2022)	✓	512	0.677	0.073	0.007
PDAE (Random $\mathbf{x}_T$) (Zhang et al., 2022)	✓	512	0.689	0.098	5.01×10^{-3}
DBAE (Stochastic) (Kim et al., 2025)	✓	512	0.920	0.094	4.81×10^{-3}
DBAE-d (Deterministic) (Kim et al., 2025)	✓	512	0.953	0.072	2.49×10^{-3}
ArcDAE (Ours)	✓	512	**0.962**	**0.065**	2.79×10^{-3}

Results & Analysis II

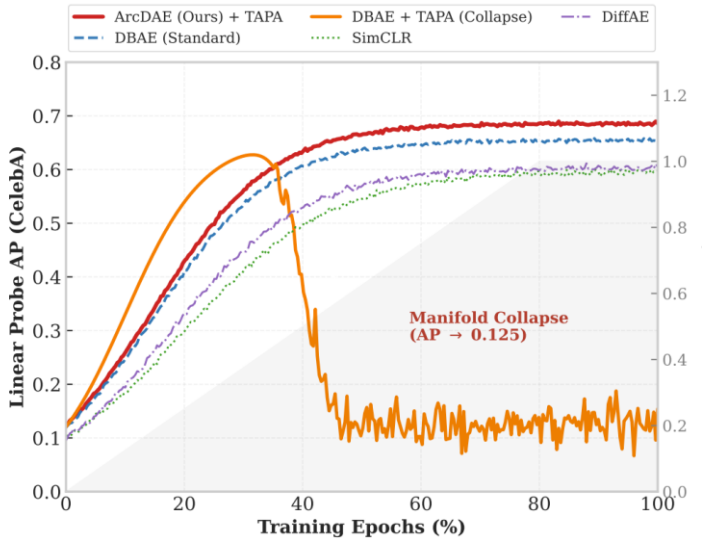
Sifted latent manifolds support generation, sample efficiency, and stability.



Method	Reg	TAD (↑)	ATTRS (↑)	FID (↓)
AE	✗	0.042±.004	1.0±.0	90.4±1.8
DiffAE (Preechakul et al., 2022)	✗	0.155±.010	2.0±.0	22.7±2.1
DBAE (Kim et al., 2025)	✗	0.124±.078	2.2±1.3	11.8±.2
ArcDAE	✗	**0.210**±.070	**3.9**±.6	**10.28**±.2
VAE (Kingma & Welling, 2014)	✓	0.000±.000	0.0±.0	94.3±2.8
β-VAE (Higgins et al., 2017)	✓	0.088±.051	1.6±.8	99.8±2.4
InfoVAE (Zhao et al., 2019)	✓	0.000±.000	0.0±.0	77.8±1.6
InfoDiff (Wang et al., 2023)	✓	0.299±.006	3.0±.0	22.3±1.2
DisDiff (Yang et al., 2023)	✓	0.305±.010	–	18.3±2.1
DBAE+TC (Kim et al., 2025)	✓	0.362±.036	3.8±.8	13.4±.2
ArcDAE+TC	✓	**0.385**±.036	**4.9**±.8	**12.2**±.2

Method	Prec (↑)	IS (↑)	FID (↓)	Rec (↑)
DDPM (Ho et al., 2020)	0.768	3.11	**9.14**	0.335
DiffAE (Preechakul et al., 2022)	0.762	2.98	9.40	**0.458**
PDAE (Zhang et al., 2022)	0.695	2.23	47.42	0.153
DBAE (Kim et al., 2025)	0.780	3.87	11.25	0.392
ArcDAE (Ours)	**0.785**	**3.92**	10.28	0.437
DiffAE + AE	0.750	3.63	2.84	0.685
DBAE + AE	0.751	3.57	1.77	0.687
ArcDAE + AE	**0.792**	**4.10**	**1.65**	**0.710**

Disentanglement and generation



Teacher anchor prevents collapse

Key takeaways

- 1% labels: ArcDAE AP 0.58 vs. DBAE 0.53.
- ArcDAE+TC reaches TAD 0.385 with FID 12.2.
- Full model gives best ablation: AP 0.681, LPIPS 0.065, FID 10.28.

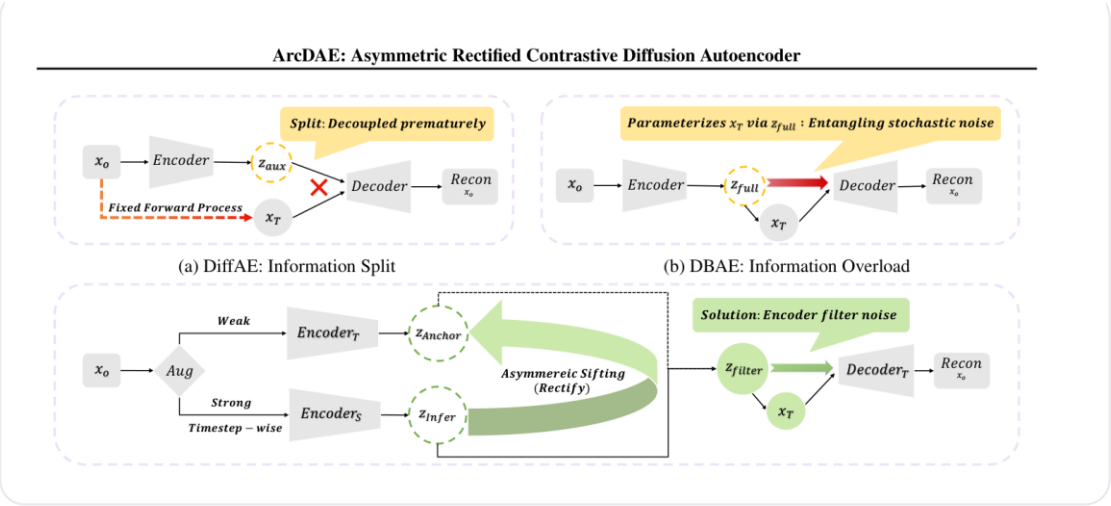
Conclusion & Future Work

ArcDAE resolves the split–overload tension with asymmetric information sifting.

Main conclusion

ArcDAE turns the diffusion bridge from a passive information carrier into an adaptive semantic sifter.

Problem	Information split loses completeness; information overload entangles noise.
Method	ASR + TAPA + TGA/TAAW rectify noisy representations toward clean semantic anchors.
Evidence	Better semantic probing, perceptual fidelity, generation quality, and ablation stability.
Future	Tighter theory for rectification dynamics; multimodal and large-scale representation learning.



Balanced bridge principle

Take-home message

Noise-rectified representations can unify synthesis and understanding.

Acknowledgement

Thank you for your attention.

ArcDAE: Asymmetric Rectified Contrastive Diffusion Autoencoder for Unified Representation Learning

Authors: Ge Gao, Di Xiong, Zeke Xie, Jian Yang, Shuo Chen*

Supported by Nanjing University and collaborating institutions.
We thank the ICML reviewers and the broader diffusion representation learning community.

Thank you!

ICML 2026