RETHINKING AUTOENCODER EVALUATION FOR CONTROLLABLE LATENT DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoencoders (AEs) define the latent space in which latent diffusion models learn the data distribution, making them a critical component of image generation. Recent AE evaluation disproportionately emphasizes whether a representation is diffusion-friendly, using image quality after diffusion training to compare and select AEs. In this work, we first formulate the condition-alignment objective of controllable latent diffusion. By decomposing the generation process and the conditional score, we derive three complementary dimensions of AE evaluation: faithfulness of condition preservation through reconstruction, recoverability of conditions from noisy latents, and tractability of learning the induced latent distribution. Existing practice prioritizes tractability, provides incomplete coverage of faithfulness through reconstruction metrics, and leaves condition recoverability unevaluated in the studies we examine. We examine representative AE methods and their variants, combining model comparisons with downstream controllable-generation experiments to answer three questions: whether the dimensions provide distinct information, whether they inform controllable-generation performance, and how to evaluate them efficiently. Our results show that the three dimensions can favor different AEs, and that condition preservation and prediction complement generation quality in assessing downstream control. Their relationship to control alignment depends on both the condition and the generator. Our recoverability experiments further show largely stable AE rankings across predictor designs and noise levels, supporting efficient comparison with a lightweight predictor evaluated on clean latents. This work broadens AE evaluation beyond diffusion-friendly image synthesis, providing a theoretically grounded and practical perspective on representations for controllable generation.

1 INTRODUCTION

An autoencoder (AE) defines the latent space in which a latent diffusion model learns the data distribution. Its encoder determines the representation available to the generator, and its decoder maps generated latents back to images. Moving generation into this representation makes high-resolution synthesis computationally practical (Rombach et al., 2022; Esser et al., 2021). The choice of AE therefore shapes both the information available for generation and the distribution the diffusion model must learn.

Recent AE research has advanced this design by studying reconstruction–generation trade-offs and developing representations that work well with diffusion models (Yao et al., 2025b; Leng et al., 2025; Zheng et al., 2025; Fan et al., 2025). Evaluation has correspondingly emphasized whether the latent space is *diffusion-friendly*, using the image quality achieved after diffusion training to compare and select AEs. Metrics such as gFID make the generation side of AE design especially visible: they summarize how effectively a representation, generator, and training recipe support realistic image synthesis.

However, as text-to-image diffusion models scale and their generation quality improves, they increasingly serve as foundations for broader applications, including spatially controlled generation, reference-guided synthesis, and image editing. These applications typically reuse the base model's AE, so the representation learned for image synthesis also shapes how image content can be preserved and controlled downstream. This work is therefore driven by a practical need: *to evaluate the AE's*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

impact on downstream controllable generation alongside the image quality it supports in the base model.

To this end, we formulate condition alignment as the expected discrepancy between the requested condition and the condition extracted from a generated image. We decompose the conditional output distribution into latent generation and decoding. The decoding component leads to *faithfulness*: how well AE reconstruction preserves each input image's condition. The resulting condition reconstruction error, which we call *condition drift*, equals the expected alignment loss under ideal conditional latent fitting; a deviation bound connects this reference to an actual generator. We then apply Bayes' rule to decompose the conditional score learned during denoising. Its condition-posterior term leads to *recoverability*, or how well conditions can be predicted from noisy latents, while fitting the full conditional latent distribution defines *tractability* for a chosen generator family. These derivations ground the three evaluation dimensions in the mathematical structure of conditional generation: what the AE preserves, what its latents reveal about the condition, and what distribution the generator must learn.

Our empirical study asks three questions: whether these dimensions provide distinct information, how they inform controllable-generation performance, and how to evaluate them efficiently. We first compare representative AE methods and their variants across generation quality, reconstruction, and condition prediction. The resulting rankings reveal complementary assessments, with paired reconstruction closely tracking condition preservation. We then connect these measurements to downstream control, finding that preservation and prediction help interpret condition alignment alongside image quality, and that their relationship depends on the control task and generator. Finally, we examine how to turn these findings into a practical evaluation. Reconstruction and generation metrics supply established measurements for faithfulness and tractability; experiments across predictor designs and noise levels support a compact recoverability comparison using a lightweight predictor on clean latents.

Together, the theory and experiments broaden AE evaluation from diffusion-friendly image synthesis to the requirements of controllable generation. They provide a connected basis for choosing what to measure, interpreting the resulting comparisons, and selecting representations for their intended downstream use.

2 Motivation: A Tilted Seesaw in Evaluation

AE design balances faithful reconstruction with a latent space that supports effective diffusion generation. Table 1 examines how eight representative AE studies report metrics and select configurations in their main papers.

Reconstruction reporting is uneven: two studies report no reconstruction metrics for their final models, four report only rFID, and two include paired metrics. Whereas rFID compares the overall reconstruction distribution, PSNR, LPIPS, and SSIM retain the correspondence between each input and its reconstruction.

AE	Reported metrics				Ablation selection	
	rFID	PSNR	LPIPS	SSIM	gFID	rFID
VA-VAE	✓	✓	✓	✓	✓	×
REPA-E	✓	×	×	×	✓	×
RAE	✓	×	×	×	✓	×
VTP	✓	×	×	×	✓	✓
SVG	✓	×	×	×	✓	×
UAE	✓	✓	✓	×	–	–
FAE	×	×	×	×	–	–
SVG-T2I	×	×	×	×	–	–

Table 1: **Evaluation reporting and ablation selection.** Checks mark reported metrics or a selected configuration with the best score; crosses mark the converse. Gray cells denote studies outside the ablation comparison.

In the five studies included in the ablation comparison, the selected configuration always attains the best gFID; only VTP simultaneously attains the best rFID. These observations reveal a *tilted evaluation seesaw*: generation quality consistently guides selection, while paired reconstruction is less systematically evaluated.

For downstream controllable generation, the AE must also support preserving and specifying image content. This motivates analyzing which AE properties these applications require, starting from a formulation of condition alignment and the AE's role in the generation process.

3 From Conditional Generation to AE Evaluation

To determine what AE evaluation should measure, we first formulate the expected condition-alignment loss of a latent diffusion model. We express the generated condition distribution through conditional latent generation and decoding, connecting the alignment objective to AE reconstruction and latent distribution fitting. We then decompose the conditional score using Bayes' rule to examine condition information within the latent representation. These derivations provide the mathematical basis for faithfulness, recoverability, and tractability.

3.1 From Generated Latents to Image Conditions

Let $(X, C) \sim p_{\mathrm{data}}$ denote an image and its condition. An encoder $\mathcal{E}$ maps X to the clean latent $Z_0 = \mathcal{E}(X)$, and a decoder $\mathcal{D}$ produces the reconstruction $\widehat{X} = \mathcal{D}(\mathcal{E}(X))$. Encoding the data induces a conditional latent distribution $p_{\mathcal{E}}(z_0 \mid c)$. A learned conditional generator instead samples from $p_\theta(z_0 \mid c)$. The forward noising process provides intermediate states Z_t on which the denoiser is trained. For conditions obtained from images, a condition extractor ϕ gives $C = \phi(X)$; examples include edge detectors, depth estimators, and segmentation models.

For a requested condition c, let $Z_0 \sim p_\theta(z_0 \mid c)$ be the generated latent and let $C' = \phi(\mathcal{D}(Z_0))$ be the condition extracted from the generated image. With the decoder and condition extractor fixed, their output depends on the generated latent. Writing the resulting probability kernel as $p_{\mathcal{D}}(c' \mid z_0)$, the distribution of conditions in generated images is

$$p_\theta(c' \mid c) = \int p_{\mathcal{D}}(c' \mid z_0)\, p_\theta(z_0 \mid c)\, \mathrm{d}z_0. \tag{1}$$

The two kernels describe consecutive stages of conditional generation. The generator samples a latent for the request; the decoder and condition extractor determine the condition present in the resulting image. For a condition-space metric $\ell_{\mathcal{C}}$, the expected alignment loss is therefore

$$\mathcal{L}_{\mathrm{align}}(\theta) = \mathbb{E}_c\, \mathbb{E}_{Z_0 \sim p_\theta(\cdot \mid c)} \left[\ell_{\mathcal{C}}\big(\phi(\mathcal{D}(Z_0)), c\big) \right]. \tag{2}$$

We begin with latents obtained by encoding real images, which provide the AE-induced target for conditional generation. Decoding these latents makes paired AE reconstruction a natural starting point for measuring condition preservation.

3.2 Faithfulness: Preserving Conditions Through Reconstruction

For $C = \phi(X)$, we compare the condition in each input image with the condition in its AE reconstruction. This condition reconstruction error is the *condition drift*; its per-image value and population mean are

$$\begin{aligned} \Delta^{\phi}_{\mathrm{AE}}(x) &= \ell_{\mathcal{C}}\big(\phi(x), \phi(\mathcal{D}(\mathcal{E}(x)))\big), \\ \Delta^{\phi}_{\mathrm{AE}} &= \mathbb{E}_{X \sim p_{\mathrm{data}}} \Delta^{\phi}_{\mathrm{AE}}(X). \end{aligned} \tag{3}$$

We call preservation of each input image's condition *faithfulness*. It concerns the encoder and decoder together, as measured on paired inputs and reconstructions.

Its connection to generation follows by substituting the AE-induced target into Eq. (2). When $p_\theta(z_0 \mid c) = p_{\mathcal{E}}(z_0 \mid c)$,

$$\begin{aligned} \mathcal{L}_{\mathrm{align}}(\theta) &= \mathbb{E}_{X \sim p_{\mathrm{data}}} \ell_{\mathcal{C}}\big(\phi(\mathcal{D}(\mathcal{E}(X))), \phi(X)\big) \\ &= \Delta^{\phi}_{\mathrm{AE}}. \end{aligned} \tag{4}$$

Condition drift therefore supplies an ideal-fit reference for controllable generation: it is the expected alignment loss when the generator learns the AE-induced conditional distribution. Measuring faithfulness identifies this reference before diffusion or controller training.

The input–reconstruction pairing also explains the relevance of reconstruction metrics. For a K_ϕ-Lipschitz condition extractor, $\Delta^\phi_{\mathrm{AE}}(x) \le K_\phi \|x - \mathcal{D}(\mathcal{E}(x))\|$. Paired reconstruction therefore constrains condition drift through the input–reconstruction relationship. Reconstruction-distribution metrics instead compare sets of images without using their pairing. For example, exchanging the reconstructions of two equally likely images leaves their image distribution unchanged while changing both paired conditions. The corresponding rFID can be zero even when condition drift is positive. Appendix A.2 gives the construction. Paired image metrics and target-condition errors thus offer connected ways to evaluate faithfulness, at general and task-specific levels.

To connect condition drift to an actual generator, define its conditional latent fitting error

$$\mathcal{W}_{\mathrm{fit}}(\theta) = \mathbb{E}_c W_1\big(p_\theta(\cdot \mid c), p_{\mathcal{E}}(\cdot \mid c)\big). \tag{5}$$

Here W_1 is the Wasserstein-1 distance in latent space.

Proposition 1 (Alignment under imperfect latent fitting)**.** *Suppose $\mathcal{D}$ is $L_{\mathcal{D}}$-Lipschitz, ϕ is K_ϕ-Lipschitz under $\ell_{\mathcal{C}}$, and the conditional latent distributions have finite first moments. Then*

$$\left|\mathcal{L}_{\mathrm{align}}(\theta) - \Delta^\phi_{\mathrm{AE}}\right| \le K_\phi L_{\mathcal{D}}\, \mathcal{W}_{\mathrm{fit}}(\theta). \tag{6}$$

For a fixed c, the function $g_c(z) = \ell_{\mathcal{C}}(\phi(\mathcal{D}(z)), c)$ is $K_\phi L_{\mathcal{D}}$-Lipschitz. Its expectation changes by at most this constant times the Wasserstein distance when the latent distribution changes. Averaging over conditions yields Eq. (6); the derivation appears in Appendix A.3.

This relation connects the two stages in Eq. (1). The AE determines the ideal-fit alignment reference, and the generator's conditional latent fitting error controls the size of the possible departure from that reference. To understand the learning problem behind this departure, we next examine the condition information and distribution that the denoiser encounters at each noise level.

3.3 Recoverability: Predicting Conditions from Noisy Latents

Faithfulness follows the condition into the reconstructed image; recoverability examines it within the representation used for generation. Consider the encoded data, $(X, C) \sim p_{\mathrm{data}}$ and $Z_0 = \mathcal{E}(X)$. The AE and forward process induce a joint distribution $p_{\mathcal{E}}(z_t, c)$ at each time t. In the score formulation, the conditional denoiser represents $\nabla_{z_t} \log p_{\mathcal{E}}(z_t \mid c)$. For positive differentiable densities at $t > 0$, Bayes' rule gives

$$\underbrace{\nabla_{z_t} \log p_{\mathcal{E}}(z_t \mid c)}_{\text{conditional score}} = \underbrace{\nabla_{z_t} \log p_{\mathcal{E}}(c \mid z_t)}_{\text{condition-posterior term}} + \underbrace{\nabla_{z_t} \log p_{\mathcal{E}}(z_t)}_{\text{marginal latent score}}. \tag{7}$$

The posterior term describes how the requested condition changes the score at a latent state. The marginal term describes the latent distribution without conditioning on a particular request. This suggests two complementary evaluations within latent generation: predicting the condition carried by a latent state and fitting the full conditional distribution of those states.

We characterize the first property through *recoverability*: how accurately the condition can be predicted from Z_t. Its Bayes risk under $\ell_{\mathcal{C}}$ is

$$\mathcal{R}^\star_t = \inf_{q_t} \mathbb{E}\, \ell_{\mathcal{C}}(C, \widehat{C}_t), \qquad \widehat{C}_t \sim q_t(\cdot \mid Z_t), \tag{8}$$

where q_t ranges over measurable predictors. This is a decision-theoretic view of the condition posterior in Eq. (7). It measures the best possible condition prediction at a given noise level. The conditional generator receives c explicitly and learns the conditional score; a condition predictor instead estimates that condition from Z_t. Condition prediction and fitting the generative distribution therefore answer separate evaluation questions.

Two simple relations connect it to the preceding analysis. At the clean endpoint, decoding followed by condition extraction is one possible prediction rule:

$$\mathcal{R}^\star_0 \le \mathbb{E}\, \ell_{\mathcal{C}}\big(C, \phi(\mathcal{D}(Z_0))\big) = \Delta^\phi_{\mathrm{AE}}. \tag{9}$$

The decoder supplies one route to the condition; another predictor may expose condition information beyond what the fixed decoder preserves. Under the forward Markov chain $C \to Z_0 \to Z_t$,

$$\mathcal{R}^\star_0 \le \mathcal{R}^\star_t. \tag{10}$$

Every prediction strategy available from Z_t can be simulated from Z_0 by applying the forward transition first. The clean state thus supports at least as accurate an optimal condition prediction as a noised state. Together, these relations connect reconstruction-based preservation to latent prediction and motivate examining prediction along the noise trajectory.

In practice, we train fixed-capacity predictors to estimate conditions from noisy latents at multiple noise levels, which we call *trajectory condition prediction* (TCP). We report task-specific prediction scores on ImageNet and compare the resulting AE rankings across predictor designs and noise levels. These scores describe condition accessibility under the specified prediction setup; Appendix B gives the implementation details.

3.4 Tractability: Fitting the Conditional Latent Distribution

The full conditional score in Eq. (7) brings condition information and latent distribution modeling together. Beyond predicting a condition from a latent state, generation requires learning how latent states are distributed for that condition. Fitting this distribution depends on both the AE and the generator family. For a conditional generator family Θ, define the best achievable fitting error

$$\mathfrak{T}_\Theta(\mathcal{E}) = \inf_{\theta \in \Theta} \mathcal{W}_{\text{fit}}(\theta). \tag{11}$$

This defines *tractability* relative to the generator family: a smaller value means that the family can fit the conditional latent distribution more closely. The full conditional score includes both the condition-posterior term and the marginal latent score in Eq. (7).

Since tractability concerns how well diffusion models learn the AE-induced latent distribution, it is typically assessed after diffusion training through generation-quality metrics such as gFID and Inception Score (IS). In practice, this dimension has already received substantial attention and is the primary focus of existing AE evaluation.

The three perspectives now specify what to examine: condition preservation through reconstruction, condition prediction from latents, and distribution learning by the generator. We next study how their assessments relate across AEs, how they inform downstream control, and how to obtain them efficiently.

4 Discussion and Analysis

Section 3 derives three perspectives from conditional generation: preserving the condition through reconstruction, recovering it from the latent representation, and learning the conditional latent distribution. We now examine what each perspective contributes to AE evaluation. The discussion first compares the information provided by their measurements, then connects these assessments to downstream control, and finally develops an efficient way to obtain them. This progression answers three research questions:

RQ1. Do the three dimensions provide distinct information? (Section 4.1)

RQ2. How do these dimensions relate to controllable-generation performance? (Section 4.2)

RQ3. How can the three dimensions be evaluated efficiently? (Section 4.3)

4.1 Relationships among the Three Evaluation Dimensions

RQ1 asks whether the three theoretically motivated perspectives lead to interchangeable AE assessments. We compare generation quality, condition preservation, and condition prediction, then examine which standard metrics already capture preservation.

We examine 33 ImageNet checkpoints from six AE families, retaining each generation score's reported recipe. Appendix B gives the measurement and training details, and Appendix D provides all checkpoint-level results and condition-wise correlations. Table 2 shows the six representative checkpoints for which reported gFID is available.

The rankings separate generation quality from condition preservation. RAE achieves the strongest reported gFID, while UAE has the lowest mean condition drift and highest PSNR. Condition prediction

	Reported gen.	Reconstruction		Faithfulness	Recoverability	
AE	gFID ↓	rFID ↓	PSNR ↑	Mean drift ↓	Canny ↑	Depth ↓
VA-VAE	2.17	0.43	24.48	0.0709	0.3064	0.1586
REPA-E	1.69	0.45	24.58	0.0700	0.4026	0.1657
RAE	**1.51**	0.77	18.41	0.1092	0.2127	**0.1448**
VTP-B	3.88	0.98	23.85	0.0836	0.2894	0.1501
SVG	3.36	0.89	21.95	0.0956	0.1768	0.1677
UAE	1.68	**0.24**	**28.26**	**0.0463**	**0.5873**	0.1675

Table 2: **Representative AE evaluation results on ImageNet.** Mean drift averages the paired condition errors for edges, depth, and segmentation. Recoverability columns report trajectory condition prediction (TCP); the arrows retain each task's score direction. The complete reconstruction comparison contains 33 checkpoints.

adds a task-specific comparison: UAE leads for Canny, whereas RAE leads for depth. Thus, both the property being evaluated and the requested condition influence the preferred AE.

Figure 1 quantifies how standard metrics relate to condition preservation. Across the six checkpoints with reported gFID, the Spearman correlation with mean condition drift is 0.086. Across all 33 checkpoints, the corresponding rFID correlation is 0.923. Paired reconstruction metrics agree still more closely with condition drift: the absolute correlations are 0.989 for PSNR and 0.953 for LPIPS. This pattern follows the measurement distinction in Section 3.2. Reconstruction realism captures much of the broad quality trend, while paired reconstruction directly retains the input–output correspondence used by condition drift. The close agreement gives paired reconstruction a practical role in evaluating faithfulness; generation quality and condition prediction contribute additional assessments.

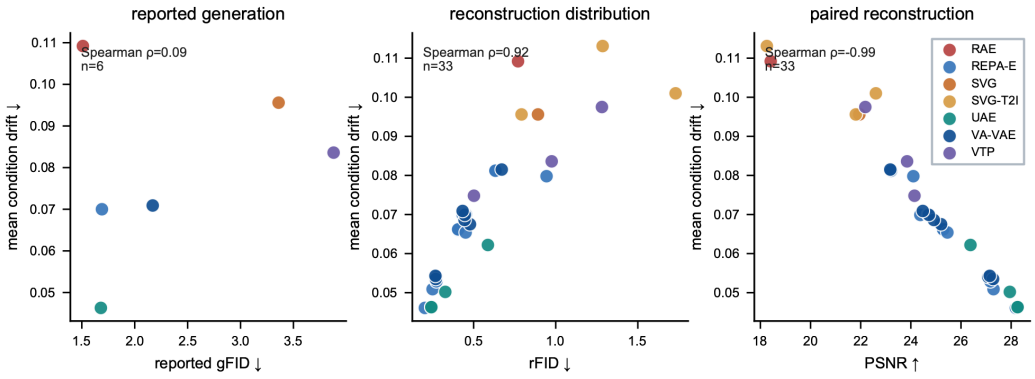

Figure 1: **How standard metrics order condition drift.** Reported gFID and mean condition drift have little rank agreement over the six available checkpoints. rFID tracks the broader reconstruction trend over 33 checkpoints, and paired PSNR gives the closest ordering. Colors identify AE families; each panel states its comparison size.

Conclusion 1. AE evaluation combines complementary perspectives

AE comparison should jointly assess generation quality, condition preservation, and condition prediction, with paired reconstruction providing a practical basis for faithfulness.

4.2 Connecting AE Evaluation to Controllable Generation

Having established how the measurements compare AEs, RQ2 asks how those assessments relate to actual control. Section 3 provides the reasoning: faithfulness gives an ideal-fit alignment reference in Eq. (4), recoverability describes the condition information associated with the posterior term

in Eq. (7), and tractability concerns learning the full conditional latent distribution. The finite-fit relation in Eq. (6) connects actual alignment to that reference through the generator's fitting error. We therefore read reconstruction and prediction measurements alongside generation quality and the condition alignment achieved by the trained system.

Table 3 follows these questions for the LDM, MAE, and DINO variants of VA-VAE, with results grouped by LightningDiT-L/2 and LightningDiT-XL/2 diffusion backbones. Each AE checkpoint is fixed and reused with both diffusion backbones, so its faithfulness and recoverability measurements are shared across the two groups. Within the L/2 group, the three variants use matched ControlNet training with shared downstream architectures and training settings. Reconstruction-time Canny F1 and depth RMSE measure faithfulness, trajectory condition prediction (TCP) measures recoverability with a specified predictor, and base/controlled gFID together with condition alignment describe generation performance.

AE	Base	Faithfulness		Recoverability		Canny ctrl.		Depth ctrl.	
	gFID ↓	Canny F1 ↑	Depth RMSE ↓	TCP-C ↑	TCP-D ↓	F1 ↑	gFID ↓	RMSE ↓	gFID ↓
LightningDiT-L/2									
LDM	6.29	**0.5453**	**0.0419**	**0.4314**	0.2390	**0.2834**	25.52	0.2055	12.78
MAE	5.57	0.5381	0.0426	0.4108	0.2200	0.2813	**20.03**	**0.1931**	**10.78**
DINO	**4.73**	0.5289	0.0429	0.3329	**0.1643**	0.2573	23.73	0.1966	13.29
LightningDiT-XL/2									
LDM	5.91	**0.5453**	**0.0419**	**0.4314**	0.2390	**0.2982**	23.84	0.1972	11.86
MAE	5.19	0.5381	0.0426	0.4108	0.2200	0.2945	**18.67**	0.1884	**9.91**
DINO	**4.38**	0.5289	0.0429	0.3329	**0.1643**	0.2719	22.11	**0.1842**	12.26

Table 3: **Condition preservation, prediction, and generation by diffusion backbone.** Bold marks the best value within each group. Both groups use the same fixed AE checkpoints and condition predictors, hence the shared faithfulness and recoverability scores. In the LightningDiT-L/2 group, the three VA-VAE variants share latent dimensionality, diffusion and control architectures, and downstream training settings. The base model for each variant is frozen while its control branch is trained. Ctrl. denotes controlled generation; TCP-C and TCP-D denote trajectory condition prediction for Canny and depth. Figure 2 expands the shared TCP measurements.

Image quality and condition alignment first give different AE preferences. DINO has the lowest base gFID in both groups, while LDM achieves the highest Canny generation F1 and DINO the lowest. MAE also improves on DINO in both Canny alignment and controlled image quality in each group. The condition-based measurements help interpret this difference between generation-centered selection and the control objective.

Canny gives a direct connection between AE condition measurements and downstream alignment. Faithfulness and recoverability both rank LDM ahead of MAE ahead of DINO, matching the generation F1 ordering in each backbone group. Here, preservation through reconstruction and accessibility to a predictor point to the same AE preference for control.

Depth reveals why the two condition measurements are complementary. LDM best preserves depth through reconstruction, while DINO has the lowest depth TCP error. This illustrates the distinction in Eq. (9): reconstruction evaluates the condition expressed by the fixed decoder, whereas prediction examines the condition information accessible in its input representation. The two measurements therefore describe different aspects of the same control task.

Relating these properties to generation also brings the generator into the comparison. MAE leads depth alignment in the L/2 group, with RMSE 0.1931 compared with DINO's 0.1966, whereas DINO leads in the XL/2 group, with 0.1842 compared with MAE's 0.1884. The strongest depth prediction thus coincides with the best depth alignment in the XL/2 results, while the L/2 results favor MAE. Because the AE checkpoints and their condition measurements are shared, the two groups show how the downstream setting participates in realizing the control task.

Generation quality completes this comparison. MAE retains the lowest controlled gFID for both tasks in both groups; in the XL/2 depth results, the alignment leader is DINO and the image-

quality leader is MAE. Faithfulness and recoverability describe the condition-related properties of the representation, while generation quality evaluates how the trained AE–generator pair supports image synthesis. Together, these measurements connect the representation to both requirements of controllable generation: image quality and condition alignment.

Conclusion 2. AE selection follows the control task and generator

AE selection should consider faithfulness, recoverability, and tractability together, with respect to the intended control task and generator.

With the complementary roles and downstream relevance established, RQ3 asks how to make this combined assessment efficient in practice.

4.3 Efficient Evaluation in Practice

Faithfulness. The close relationship in Section 4.1 makes paired reconstruction a natural starting point for faithfulness evaluation. Paired metrics such as PSNR, SSIM, and LPIPS describe reconstruction fidelity, while applying the target condition extractor to both images gives a task-specific preservation score. This combines a familiar reconstruction assessment with direct measurement of the information needed for the intended control task.

Tractability. Tractability is assessed after diffusion training through generation metrics such as gFID and Inception Score. Using a common generator architecture and training recipe gives a consistent basis for comparing how effectively candidate latent representations support image synthesis in the intended setting.

Recoverability. Recoverability adds a condition-prediction task, so we ask how lightweight this additional measurement can be. We examine whether the relative AE assessment remains stable when the predictor design and evaluation noise change. Figure 2 compares Base, Base (wider), Base (wider + residual), and Base (shallower) on the three matched AE models over $t \in \{0, 0.1, \ldots, 0.9\}$.

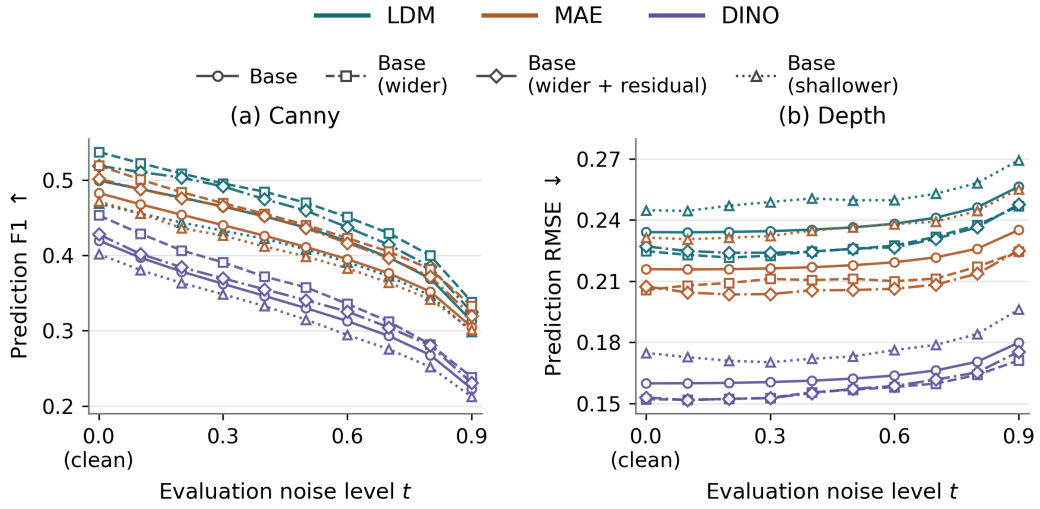

Figure 2: **Condition prediction across AE and predictor variants.** Colors distinguish the fixed VA-VAE variants; markers and line styles distinguish the four TCP predictor designs. Each panel contains twelve curves evaluated on the same ten-point noise grid, including clean latents at $t = 0$. Base supplies the TCP measurements used in Table 3.

Predictor design changes the absolute scores, while all four designs give the same ten-level mean rankings: LDM ahead of MAE ahead of DINO for Canny, and DINO ahead of MAE ahead of LDM for depth. Depth preserves its ordering in all 40 predictor–noise comparisons. Canny retains the same ordering in 37 of 40; the other three are small LDM/MAE swaps under the shallower predictor at

$t = 0, 0.1, 0.9$, with F1 differences below 0.003. The lightweight, trajectory-trained Base predictor at $t = 0$ recovers the mean ordering shared by all four designs. This gives a compact evaluation choice: compare the AEs with a common lightweight predictor evaluated on clean latents. Table 3 retains the Base TCP measurements, and Appendix B.2 gives the Base protocol and predictor-design comparison.

Conclusion 3. Lightweight predictors enable practical AE comparison

For these AE variants, a shared lightweight predictor evaluated on clean latents provides an efficient comparison of recoverability.

5 Related Work

Autoencoders for generative modeling. Two-stage generative models separate image tokenization from learning a prior over the resulting representation (van den Oord et al., 2017; Razavi et al., 2019; Esser et al., 2021; Chang et al., 2022). Latent diffusion extends this design to efficient high-resolution synthesis (Rombach et al., 2022; Peebles & Xie, 2023). Recent AE work studies reconstruction–generation trade-offs, end-to-end adaptation, representation encoders, and more compact visual tokens (Yao et al., 2025b; Leng et al., 2025; Zheng et al., 2025; Yao et al., 2025a; Shi et al., 2025b; Fan et al., 2025; Gao et al., 2025; Yu et al., 2024).

Latent representations and condition prediction. SVG combines frozen self-supervised features with a residual branch to retain semantic discriminability and reconstruction detail (Shi et al., 2025b). Geometric Autoencoder (GAE) studies semantic supervision, latent normalization, and dynamic noise sampling to shape the representation used by diffusion (Liu et al., 2026). These designs connect representation content with the learning problem faced by diffusion. Our recoverability perspective makes the condition-specific part of this connection measurable through prediction from latent states. The predictor and noise-level comparisons then examine how this measurement can be used efficiently for AE comparison.

Control and editing in diffusion models. ControlNet and related adapters introduce spatial and reference conditions into pretrained generators (Zhang et al., 2023; Mou et al., 2024; Li et al., 2023; Ye et al., 2023; Zhao et al., 2023; Huang et al., 2023; Li et al., 2024). Image editing and personalization further emphasize instance-specific preservation (Meng et al., 2022; Brooks et al., 2023; Gal et al., 2023; Ruiz et al., 2023). These tasks make both image quality and agreement with an individual request central to evaluation. We connect those outcomes to the AE through condition preservation and prediction measured directly from the fixed representation and decoder. The downstream experiments examine how these representation-level measurements relate to the behavior of the trained AE–generator pair.

Distributional and paired evaluation. FID, precision–recall, and related measures characterize the realism and coverage of image distributions (Heusel et al., 2017; Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020). PSNR, SSIM, and LPIPS instead retain a correspondence between inputs and reconstructions (Wang et al., 2004; Zhang et al., 2018). This distinction explains why reconstruction quality provides several kinds of evidence about faithfulness. Our analysis connects paired condition preservation to ideal conditional generation, and the empirical study shows how standard reconstruction metrics track this property. Together with condition prediction and generation quality, these measurements support the three-perspective evaluation developed in this work.

6 Conclusion

AE evaluation for controllable latent diffusion benefits from three complementary perspectives: faithfulness, recoverability, and tractability. Our analysis connects them to conditional generation, and our experiments show how their assessments relate to one another and to downstream control. Combining paired reconstruction, generation quality, and lightweight condition prediction turns these perspectives into a practical evaluation. This broadens AE selection from image synthesis quality to the representation requirements of the intended control task and generator.

AI USE STATEMENT

We used GPT-5.6 Sol and GPT-6 Astra for the following purposes: (1) revising and polishing manuscript drafts; (2) assisting with LaTeX formatting, including generating Python scripts to convert experimental JSON outputs into LaTeX tables, such as Tables 2–5; and (3) conducting preliminary exploration of related work. All AI-assisted edits, scripts, generated tables, and literature findings were subsequently reviewed and verified by the authors before being incorporated into the manuscript.

REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, and Ziwei Liu. The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding. *arXiv preprint arXiv:2512.19693*, 2025.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Yuan Gao, Chen Chen, Tianrong Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation. *arXiv preprint arXiv:2512.07829*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *International Conference on Machine Learning*, pp. 13753–13773. PMLR, 2023.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pp. 129–147. Springer, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

Hangyu Liu, Jianyong Wang, and Yutao Sun. Geometric autoencoder for diffusion models. *arXiv preprint arXiv:2603.10365*, 2026.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pp. 7176–7185. PMLR, 2020.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Mehdi S M Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Minglei Shi, Haolin Wang, Borui Zhang, Wenzhao Zheng, Bohan Zeng, Ziyang Yuan, Xiaoshi Wu, Yuanxing Zhang, Huan Yang, Xintao Wang, Pengfei Wan, Kun Gai, Jie Zhou, and Jiwen Lu. Svg-t2i: Scaling up text-to-image latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2512.11749*, 2025a.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2510.15301*, 2025b.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. *arXiv preprint arXiv:2512.13687*, 2025a.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025b.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11127–11150, 2023.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.