A UNIFIED VIEW OF JOINTLY EVOLVING REPRESENTATIONS AND IMAGES IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models increasingly use pre-trained visual representations to guide image generation, enhancing convergence and generation quality. Existing methods, however, mostly couple representations and images at two endpoint timesteps: RCG first synthesizes a clean representation at $t = 0$ and then uses it as a static condition, while ReDi and REG jointly denoise the representation and image from pure noise at $t = T$. In this paper, we provide a unified view of understanding such contradictions and show that *both endpoints are suboptimal*. Intuitively, the former prevents representations from receiving image context, while the latter forces image denoising to rely on highly noisy representations. We formalize this trade-off as a minimization problem, where these two setbacks combined, coincide with the divergence between actual diffusion denoising and a hypothetically ideal one, with clean representations or images as conditions. Empirical probing and theoretical analysis show that the optimum lies at neither extreme endpoint, but at an intermediate coupling timestep $t = s$. Motivated by this insight, we introduce **JERI** (Jointly Evolving Representations and Images), which first denoises representations to this intermediate state and then jointly evolves representations and images. Experiments show that JERI outperforms the extreme designs in prior works across conditional and unconditional generation, model scales, and representation spaces.

1 INTRODUCTION

Pre-trained visual representations have begun to reshape image generation. Features learned by self-supervised encoders such as MoCov3 (Chen et al., 2021) and DINOv2 (Oquab et al., 2023) provide compact semantic structure beyond labels or pixels, and recent diffusion models have shown that injecting such representations can accelerate convergence and improve generation quality (Li et al., 2024a; Yu et al., 2025; Kouzelis et al., 2025; Wu et al., 2025b; Baade et al., 2026). This trend suggests a shift in the role of representations: they are no longer only descriptors extracted after generation, but can serve as generative pivots that guide, condition, and co-evolve with images. Once representations become part of the diffusion process, a basic design question arises: *whether* and *where* along the denoising trajectory should representations and images be coupled?

Interestingly, existing work answers this question in contradictory ways at the two extremes of the diffusion process. As illustrated in Fig. 1, RCG (Li et al., 2024a) first synthesizes a clean representation z_0 and then uses it as a static condition for image generation. In terms of the diffusion trajectory, this corresponds to no joint evolution of representations and images and corresponds to the endpoint $t = 0$, after the representation has already been fully denoised. In contrast, ReDi (Kouzelis et al., 2025) and REG (Wu et al., 2025b) couple representation and image entirely, from the other endpoint $t = T$, where both variables start from pure noise and are jointly denoised throughout the reverse process. In fact, these two designs show distinct emphases: RCG highlights the benefits of a clean representation for guiding image generation, while the other emphasizes the flexibility of fully joint evolution from the initial time $t = T$.

In this paper, we provide a unified view of these seemingly contradictory designs. We show that they are two endpoints of a single axis: the timestep at which representation and image begin to interact. This view also reveals the insights into why both endpoints are suboptimal. **(1)** When coupling occurs only at $t = 0$, the representation can provide clean conditioning but is generated without access to

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.

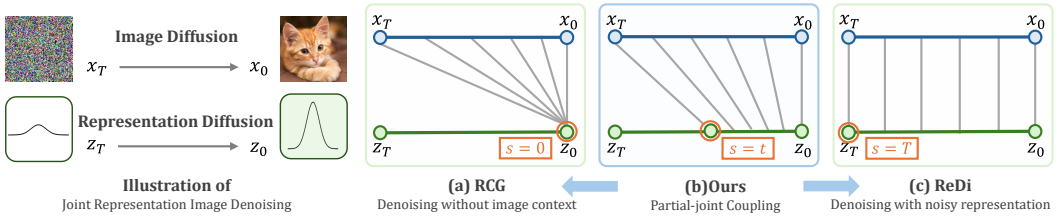


Figure 1: Representation-image coupling axis. **Left:** RCG uses a clean but static representation, giving the representation branch no image context. **Right:** ReDi couples from pure noise, so early image steps depend on noisy representations. **Center:** JERI shows that the optimal coupling timestep is strictly within the diffusion process with theoretical and empirical analysis.

the images, which cannot adapt to and align with such posterior information. **(2)** On the other hand, when coupling starts from $t = T$, the representation and image have the flexibility to jointly evolve, but the image branch must condition on a representation that is still highly noisy, especially in the early denoising process. Based on the formulation of diffusion models, such noisy representations z_t can only provide unreliable guidance with low mutual information with a clean representation z_0.

These two endpoints form our basic intuition on the trade-offs of coupling representations with images. We formalize this intuition by introducing a trajectory cost indexed by the coupling timestep s. Before coupling, the representation branch incurs an *imageless cost*, corresponding to denoising representations without image context. After coupling, the image branch incurs a *noisy-representation cost*, corresponding to denoising images without access to a clean representation. The total cost of a coupling schedule is obtained by accumulating these two costs along the trajectory.

This trajectory view, with clearly defined divergence between the actual diffusion trajectories and an ideal intended one, can be both empirically estimated and theoretically analyzed. Notably, we discover that the cost of choosing a coupling timestep s follows the trend as Fig. 2. This leads to a simple but important conclusion: the optimal coupling timestep does not lie at either endpoint, but strictly an intermediate timestep that could be probed.

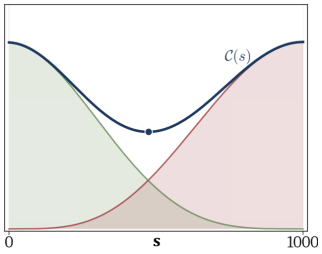


Figure 2: The green area represents *imageless cost*, and pink area shows *noisy-representation cost*, and their sum $\mathcal{C}_{\text{JERI}}(s)$ is the minimization target. Our probing shows its optimum at the intermediate timestep (details in Sec. 4).

Motivated by this analysis, we introduce **JERI**, a partially joint framework for representation-image diffusion, emphasizing the importance of selecting an intermediate coupling timestep. Concretely, JERI first denoises the representation independently to an intermediate state z_s, and then jointly evolves representations and images through lightweight conditioning both from images to representations and representations to images. The representation-to-image pathway allows the current representation state to guide image denoising, while the image-to-representation pathway allows the forming image to refine the representation trajectory. This design approaches the interior coupling suggested by our trajectory-cost analysis: the representation becomes sufficiently informative before it guides image generation, while later representation states remain adaptive to the generated image.

We instantiate JERI on top of DiT-based image diffusion and an MLP-based representation diffusion branch. To stabilize the partially joint process, we further adopt teacher forcing with stop-gradient, which keeps each branch grounded in its own diffusion objective. We evaluate JERI across both conditional and unconditional generation under varied scales. Across all settings, JERI consistently outperforms prior work built on extreme designs.

Our contributions are summarized as follows:

1. We provide a unified view of representation-image diffusion, showing that RCG and ReDi/REG correspond to two endpoint coupling schedules along the same timestep axis.
2. We identify two complementary information costs behind these endpoints: the *imageless cost* of denoising representations without image context, and the *noisy-representation cost* of denoising images without clean representations.
3. We formulate a trajectory-cost objective over the coupling timestep and show, through empirical probing and theoretical analysis, that the optimal coupling occurs at an intermediate timestep.

4. We introduce JERI, a partially joint representation-image diffusion framework that realizes this interior coupling schedule, achieving consistent improvements across conditional and unconditional generation, model scales, representation spaces, and data diversities.

2 Related Work

Diffusion Models for Image Generation. Diffusion models are crucial ways for visual generation (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Kingma et al., 2021; Karras et al., 2022; 2024; Rombach et al., 2022; Wu et al., 2025a). They learn a denoising process that gradually converts noisy signals into clean visual outputs and can be guided during sampling to better approximate conditional distributions (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). Recent advances further improve their scalability through stronger architectures and latent formulations, including latent diffusion models (Rombach et al., 2022), diffusion transformers (Peebles & Xie, 2023), rectified-flow transformers (Esser et al., 2024), and pixel-space transformer variants (Li & He, 2025). These developments show that diffusion models are not only powerful generative models but also flexible frameworks whose internal variables, architectures, and denoising trajectories can be redesigned for better generation.

Representation in Generation. Recent studies have revealed the pivotal role of pre-trained representations in visual generation. The REPA (Yu et al., 2025; Wang et al., 2025c; Leng et al., 2025; Wang et al., 2025b; Kouzelis et al., 2025; Wang et al., 2025a; Wu et al., 2025b) line of work utilizes pre-trained representations, *e.g.*, DINOv2 (Oquab et al., 2023), as auxiliary guidance for the intermediate features of DiT (Peebles & Xie, 2023). These explorations also encourage the decoupling and enhancement of generative models' internal representations (Jiang et al., 2025; Wang & He, 2025). The RCG (Li et al., 2024a; Lavoie et al., 2025; Zheng et al., 2025) line instead uses pre-trained representations, *e.g.*, MoCov3 (Chen et al., 2021), as a proxy or conditioning signal for image synthesis. More recent representation-autoencoder and tokenizer designs further suggest that the structure of the latent space itself is a central design factor for generation (Zheng et al., 2025; Tong et al., 2026; Yu et al., 2026; Yao et al., 2025; Chen et al., 2025c;b;a; Xiong et al., 2025; Li et al., 2024b;c; Qiu et al., 2025; Shi et al., 2025). Together, these works show a growing consensus that representations are not merely byproducts of generative training, but useful design axes for improving sample quality, scalability, and controllability.

A related line of recent work studies how different generative signals should be ordered or scheduled during generation (Baade et al., 2026; Du, 2026; Huang et al., 2025). While these methods also challenge the view that all variables should evolve symmetrically across timesteps, their schedules are mainly designed around latent-pixel ordering, frequency structure, or autoregressive consistency. In contrast, our work focuses on the coherence between pre-trained representations and images during diffusion, particularly how this relates to the divergence between the real and ideal diffusion denoising trajectories, which enables our JERI to investigate the optimal ranges of coupling schedules.

3 Preliminary

3.1 Diffusion Models in Image Generation

We follow the settings of DDPM (Ho et al., 2020) and latent diffusion (Rombach et al., 2022) for image generation. The diffusion model's forward process adds noise to an image x_0 (*e.g.*, the latents encoded by a VAE (Kingma & Welling, 2014; Rombach et al., 2022)) over T steps until reaching a Gaussian noise: $t \in \{1, ..., T\}$, $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t$ is the cumulative noise schedule. We denote the resulting Markov chain as $q(x_t|x_{t-1})$. The diffusion model then learns the backward denoising process that maps a noisier x_t into a cleaner x_{t-1}, denoted as $p(x_{t-1}|x_t)$. In this way, the diffusion models can gradually recover an image x_0 from a random Gaussian noise at x_T.

3.2 Representation in Image Generation

Existing diffusion frameworks couple representation and image generation in two extreme ways. Define an image's representation z_0 as the output of a recognition encoder, *e.g.*, DINOv2.

RCG (Li et al., 2024a)'s coupling happens at $t = 0$: a representation diffusion model first learns $p(z_{t-1}|z_t)$ to generate z_0, then an image model learns $p(x_{t-1}|x_t, z_0)$ conditioned on the clean representation z_0.

$$z_T \to \cdots \to z_0, \qquad (x_T, z_0) \quad \to \cdots \to (x_0, z_0). \tag{1}$$

As shown in Eqn. 1, z_0 is generated and then held constant, acting as a *static anchor* for the entire image synthesis process without being able to receive contexts from the image.

ReDi (Kouzelis et al., 2025) and REG (Wu et al., 2025b) couple representation and image generation from the pure noise, learning the joint backward transition $p(x_{t-1}, z_{t-1} \mid x_t, z_t)$ at every step:

$$(x_T, z_T) \to \cdots \to (x_t, z_t) \to \cdots \to (x_0, z_0). \tag{2}$$

As shown in Eqn. 2, the image always conditions on the *noisy* state z_t rather than the clean z_0.

4 Analyzing When To Couple Representations and Images

Given the two extremes we observed in Sec. 3.2, this section establishes a framework to unify them and analyze the optimal way to jointly evolve representations and images. Specifically, Sec. 4.1 and 4.2 show that each extreme of RCG and ReDi/REG exhibits a strictly positive gap compared to the ideal backward transition. Then, these two individual costs are merged and analyzed jointly in Sec. 4.3. Finally, Sec. 4.4 gives sufficient conditions for interior coupling and probes trajectory cost.

4.1 Imageless Cost

RCG's representation branch denoises in isolation from the image. At every step, it sees only the noisy representation z_t as input. Its backward transition is $p(z_{t-1} \mid z_t)$, without the image condition. We show that this modeling induces an information cost that adding model capacity cannot recover. Specifically, we consider a diffusion process as illustrated in Fig. 3, where an intermediate x_t could lead to different clean images x_0. Therefore, a natural implication of this observation is that such a static z_0 might be misaligned with the image x_t during the denoising process, as it cannot receive the contexts from the evolving images. To illustrate the gap caused by this design, we compare two backward transitions: **(a)** the joint transition $p(z_{t-1} \mid x_t, z_t)$, which uses both noisy variables; **(b)** the z-only transition $p(z_{t-1} \mid z_t)$ of RCG. The two agree exactly when x_t contributes no information about z_{t-1} beyond what z_t already supplies. The resulting gap, which we term as the *Imageless Cost*, is formalized below.

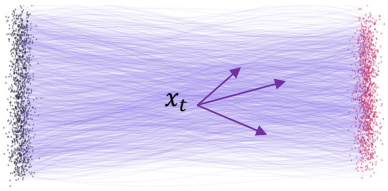


Figure 3: Dynamical nature of diffusion models. An early noisy state x_t can evolve toward multiple x_0, making a single static representation misaligned and problematic.

Proposition 4.1 (Imageless cost is strictly positive)

Suppose $\mathrm{Cov}(x_0, z_0) \neq \mathbf{0}$ *and the forward processes for* x *and* z *have independent noise. Then for every* $t \in (0, T]$,

$$p(z_{t-1} \mid x_t, z_t) \neq p(z_{t-1} \mid z_t), \quad \textit{equivalently} \quad I(z_{t-1}; x_t \mid z_t) > 0. \tag{3}$$

(Proof in Sec. A.6). With the above proposition, we denote $I_t^z := I(z_{t-1}; x_t \mid z_t)$. By Proposition 4.1, $I_t^z > 0$ on $(0, T]$ whenever x_0 and z_0 are correlated. On the whole denoising trajectory, the sum of such imageless cost for RCG is $\sum_{t=1}^{T} I_t^z$.

4.2 Noisy Representation Cost

ReDi and REG's image branch denoises with the noisy representation z_t at every step: the backward transition is $p(x_{t-1} \mid x_t, z_t)$, without the clean representation z_0. Similar to the imageless cost in RCG, this design mirrors a *noisy-representation cost*. Intuitively, the clean representation z_0 is the most informative input for generating x_0; as the noisy state z_t comes from adding Gaussian noise to z_0, an image branch utilizing z_t operates with strictly less information than using z_0.

Specifically, we show this via the divergence between two backward transitions: **(a)** the denoising process can utilize clean representations, the backward is $p(x_{t-1} \mid x_t, z_t, z_0)$; **(b)** in actual ReDi and REG, which is restricted to using noisy z_t, the backward is $p(x_{t-1} \mid x_t, z_t)$. Formally:

Proposition 4.2 (Noisy representation cost is strictly positive)

Suppose $\mathrm{Cov}(x_0, z_0) \neq \mathbf{0}$ *and the forward processes for* x *and* z *have independent noise. Then for every* $t \in (0, T]$,

$$p(x_{t-1} \mid x_t, z_t, z_0) \neq p(x_{t-1} \mid x_t, z_t), \quad \textit{equivalently} \quad I(x_{t-1}; z_0 \mid x_t, z_t) > 0. \tag{4}$$

(Proof in Sec. A.6). Similar to the imageless cost, we denote $I_t^x := I(x_{t-1}; z_0 \mid x_t, z_t)$. By Proposition 4.2, $I_t^x > 0$ on $(0, T]$ whenever x_0 and z_0 are correlated. This cost is also determined by the data distribution and the schedule formulation in diffusion models. As such a cost can lead to drifted denoising of image at every step, the total accumulated cost over the trajectory is $\sum_{t=1}^{T} I_t^x$.

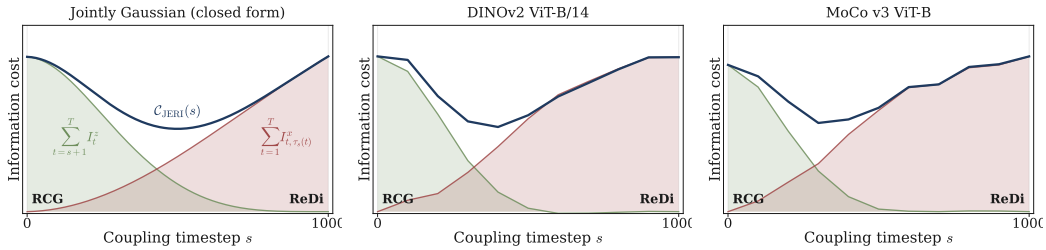


Figure 4: Probing the optimal trajectory cost $\mathcal{C}_{\text{JERI}}(s)$: decreasing imageless cost $\sum_{t=s+1}^{T} I_t^z$ and increasing noisy-representation cost $\sum_{t=1}^{T} I^x_{t,\tau_s(t)}$. All three show interior minima near the center, confirming our partial-joint design. (Left) Synthetic Gaussian diffusion. (Center, Right) ImageNet class probing.

4.3 Partial Joint Coupling and Trajectory Cost

This section defines the coupling axis between representation and image denoising. A coupling timestep s determines when the two branches begin exchanging information, with $s = 0$ giving RCG and $s = T$ giving ReDi/REG.

Propositions 4.1 and 4.2 show that both endpoint schedules incur a strictly positive gap between the ideal and available reverse transitions. We therefore unify them by allowing a coupling timestep $s \in \{0, \dots, T\}$. For $t > s$, the representation branch denoises without image context. After coupling, the image branch denoises from x_T to x_0, while the representation branch continues from z_s to z_0. We pair image timestep t with representation timestep $\tau_s(t) := \lfloor \frac{s}{T} t \rfloor$, so that both branches reach clean states simultaneously. The trajectory is

$$z_T \to \cdots \to z_s, \qquad (x_T, z_s) \to \cdots \to (x_t, z_{\tau_s(t)}) \to \cdots \to (x_0, z_0). \tag{5}$$

Thus Eq. (5) interpolates between the two endpoints: $s = 0$ gives RCG, and $s = T$ gives ReDi/REG.

Trajectory cost. To find a favorable coupling timestep s, we measure the information gap between the oracle reverse transition and the reverse transition available under the coupling schedule. Along the trajectory of Eq. 5, the representation branch incurs the Imageless Cost I_t^z while it denoises without image context before coupling, and the image branch incurs the Noisy Representation Cost when it is conditioned on the paired representation state $z_{\tau_s(t)}$ rather than clean z_0. Aggregating these two costs gives the trajectory cost

$$\mathcal{C}_{\text{JERI}}(s) := \sum_{t=s+1}^{T} I_t^z + \sum_{t=1}^{T} I^x_{t,\tau_s(t)}. \tag{6}$$

Here $I^x_{t,\tau}$ denotes the image-branch cost conditioned on z_τ, so $I^x_{t,t} = I^x_t$ and $I^x_{t,0} = I(x_{t-1}; z_0 \mid x_t, z_0) = 0$. At the endpoints, $\mathcal{C}_{\text{JERI}}(0) = \sum_{t=1}^{T} I_t^z$ recovers the full imageless cost of RCG, while $\mathcal{C}_{\text{JERI}}(T) = \sum_{t=1}^{T} I_t^x$ recovers the full noisy-representation cost of ReDi/REG. Appendix A.10 shows that the same information terms measure the errors of finite-step sampling, linking trajectory cost to the generated images.

4.4 Analyzing Trajectory Cost: An Intermediate Timestep is Optimal

Leaving either endpoint trades one cost for another. Moving away from $s{=}0$ gives the representation branch image context earlier but makes the image branch depend on a noisy representation. Moving away from $s{=}T$ gives the image branch a cleaner representation but delays image context for the representation branch. A coupling timestep is favorable when both trades yield a net reduction.

Formal justification. The theorem below gives sufficient conditions directly for $\mathcal{C}_{\text{JERI}}$ (Appendix A.9). For each endpoint, the proof constructs a margin of the form "*cost removed by leaving the endpoint*" minus "*cost newly introduced by leaving it.*" The actual JERI trajectory also uses the warped pairing $z_{\tau_s(t)}$, so these margins must cover the corresponding pairing difference.

$$m_0(s) := I(x_s; z_0 \mid z_s) - \big[I(x_0; z_0 \mid z_s) - I(x_s; z_0 \mid z_s)\big], \quad \Delta(s) := \sum_{t=1}^{T} I^x_{t,\tau_s(t)} - \sum_{t=1}^{s} I^x_t \tag{7}$$
$$m_T(s) := \big[I(x_s; z_0 \mid z_s) - I(x_T; z_0 \mid z_s)\big] - I(x_s; z_s \mid z_T),$$

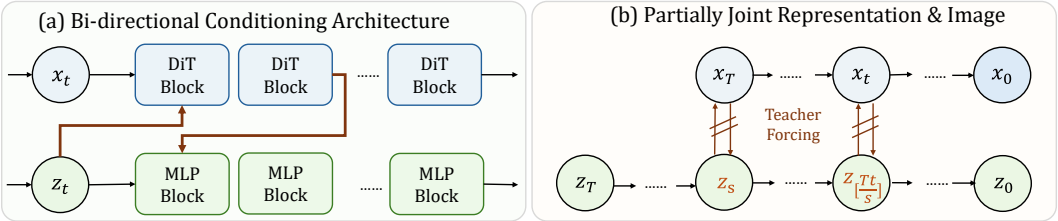


Figure 5: Overview of our JERI. **(a)** We implemented JERI with a minimal architectural change to RCG via utilizing several small MLP blocks to pass representations as conditions for image denoising, and images as conditions for representation denoising. **(b)** Building on our analysis, JERI leverages the partial-joint coupling to independently denoise representation until $t = s$, and then jointly synthesize both. *Teacher forcing* with stop gradient is applied to stabilize training.

Theorem 4.1 (Sufficient Conditions for Interior Coupling)
Assume the forward processes above and that the mutual-information terms below are finite. Suppose there exist $s_0, s_T \in \{1, \dots, T-1\}$ *such that* (C1) $m_0(s_0) - \Delta(s_0) > 0$, (C2) $m_T(s_T) - \Delta(s_T) > 0$. *Then* $\mathcal{C}_{\text{JERI}}(s_0) < \mathcal{C}_{\text{JERI}}(0), \mathcal{C}_{\text{JERI}}(s_T) < \mathcal{C}_{\text{JERI}}(T)$. *Hence every global minimizer of* $\mathcal{C}_{\text{JERI}}$ *lies in* $\{1, \dots, T-1\}$.

Probing with Synthetic Gaussian Examples. For scalar jointly Gaussian (x_0, z_0) with ρ=0.5, the conditions and $\mathcal{C}_{\text{JERI}}(s)$ have closed forms. Fig. 4 (left) shows that the interior cost falls below both endpoints: the imageless costs and noisy-representation costs follow the reverse pattern, and their sum has the minimum at approximately half of the diffusion schedule. Under the Gaussian setting, Appendix A.10 further evaluates final image KL for standard fixed-variance DDPM and Heun probability-flow sampling where both process attain interior minima.

Probing $\mathcal{C}_{\text{JERI}}(s)$ on ImageNet. We then move on to analyzing $\mathcal{C}_{\text{JERI}}(s)$ in real-world datasets like ImageNet (Deng et al., 2009), utilizing the pretrained representation space of both DINOv2 and MoCo v3. As the exact estimation of $\mathcal{C}_{\text{JERI}}(s)$ in high dimensions is intractable, we utilize class probing as a tractable diagnostic: we train two linear classifiers (Alain & Bengio, 2018), one on the clean x_0 and one on the clean z_0; once frozen, each class c turns x_t and z_t into its two scalar logits. Since a logit is a linear function, the logit of x_t is the logit of x_0 carried through the same forward process, and likewise for z. We therefore probe the Imageless Cost I_t^z and the Noisy Representation Cost $I_{t,\tau}^x$ through differences of mutual informations between these scalars, averaged over the 1000 classes (Appendix B). As shown in Fig. 4, the probing results show trends similar to our Gaussian example: both DINOv2 and MoCo v3 exhibit a clear interior minimum, strictly below both endpoints.

5 Method: JERI

5.1 Partial-Joint Evolution of Representations and Images

The analysis in Sec. 4.3 suggests a simple and general design principle: representations and images should not necessarily be coupled throughout the entire denoising trajectory, or not coupled at all. Instead, the coupling should begin from an intermediate timestep as in Eqn. 5. With the probing results in Fig. 4 being close to the midpoint of the schedule, we empirically select the timestep to be $s = T/2$, balancing the proximity to our estimation and simplicity for implementations. We ablate other choices of this timestep in Secs. 6.3–6.4.

Our schedule is architecture-agnostic: the same principle applies to unconditional generation and class-conditional generation, or any setting where an image trajectory is guided by an auxiliary representation trajectory. We choose to instantiate this principle on top of RCG (Li et al., 2024a), which provides a clean and principled architecture for us to experiment with the coupling representations and images. An overview of our method is in Fig. 5.

5.2 Bi-directional Conditioning

To instantiate the joint evolution in RCG, we propose the principled architecture as in Fig. 5a, consisting of **(1)** an image generation branch (a DiT) and **(2)** a representation synthesis branch (a series of MLP blocks). These two information pathways enable the bi-directional information exchange between the representations and images.

Representation → Image. The dynamic representation z_t is fed into the DiT blocks (similar to a class embedding) to guide image synthesis. In this way, image generation receives evolving guidance from the representation.

Image → Representation. To enable the injection of image context into representation generation, we are inspired by related studies in REPA (Yu et al., 2025) and DDT (Wang et al., 2025b), where intermediate features from the DiT blocks can be aligned with pre-trained feature spaces. Therefore, we project the features of intermediate DiT blocks into the representation generation branch, allowing the image state x_t to guide the refinement of the representation (Fig. 5a).

Loss Function. JERI does not introduce additional complexity in terms of the diffusion loss. The final optimization objective is a weighted combination of the image diffusion loss and representation diffusion loss. More details are covered in the supplementary materials.

5.3 Stabilization via Teacher Forcing

In addition to optimizing the total trajectory cost with partial joint coupling, we further notice the low alignment between representations and images, especially in earlier training steps. Instead of letting the two representation and image generation branches interfere with each other, we stabilize the optimization by detaching the gradient from the bi-directional conditioning so that each branch maintains its singular learning objective. As each branch still accepts input from the other branch as additional context, *e.g.*, the image generation DiT receives the representation z_t as guidance, our solution is similar to teacher forcing in autoregressive models. With this additional stabilization approach, our JERI learns to comprehend context from the other branch without sacrificing optimization in either image or representation generation.

6 Experiments

6.1 Implementation Details

Datasets. Following the prior work, we mainly adopt ImageNet (Deng et al., 2009). However, due to the large amount of computation required for full training on ImageNet, we (1) utilize the full ImageNet and DINOv2 feature space, training for 400k steps for a full comparison, matching the generation and evaluation settings in ReDi (Kouzelis et al., 2025); and (2) adopt a random 100-class ImageNet subset in `timm` (timm), and the MoCov3 feature space aligned with RCG, for extensive ablation studies and analysis.

Evaluation Metrics. Following existing image generation studies, we adopt the Fréchet Inception Distance (FID) (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016). To further analyze the enhancements of JERI for representation, we evaluate the FID of synthesized representations, *i.e.*, the distance between the embeddings from real images and generated representations. We term this metric *representation FID* (R-FID). Across all evaluations, we sample the same number of images as the reference image set for FID calculation as in prior works.

Architecture, Training, and Inference. **(1)** We follow RCG (Li et al., 2024a) and utilize DiT (Peebles & Xie, 2023) as our main architecture. **(2)** Our training follows the hyperparameters in DiT (Peebles & Xie, 2023), including a batch size of 256, AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2017) optimizer, a constant learning rate of 1×10^{-4}, and an exponential moving average (EMA) with a decay of 0.9999. For training on the 100-class ImageNet subset, we train all models for 600 epochs. **(3)** We follow DDPM (Ho et al., 2020) and ADM (Dhariwal & Nichol, 2021) with $T = 1000$ timesteps and respaced DDPM (SDE) sampling. Following RCG (Li et al., 2024a), both branches use fixed reverse variance $\sigma_t^2 = \beta_t$, as analyzed in Fig. A.1. More implementation details are in Sec. D.

6.2 Main Results

6.2.1 Class-Conditional Generation

We evaluate JERI on class-conditional generation as Table 1. For a fair comparison, we evaluate JERI without class-label classifier-free guidance to directly compare with the performance of the methods under the same setting, such as ReDi (Kouzelis et al., 2025). According to Table 1, JERI achieves substantially better image fidelity under the same compute across different scales. This is consistent with our motivation: through partial joint coupling, JERI avoids the full noisy-representation cost of ReDi, where early image steps condition on a

Table 1: Comparison on class-conditional generation. All methods were trained with 400k iterations and evaluated without class-label classifier-free guidance. ↓ indicates lower is better, ↑ indicates higher is better. Green text shows the FID improvement of JERI over the strongest baseline ReDi.

Model		FID ↓	IS ↑
DiT-B	DiT (Peebles & Xie, 2023)	43.5	–
	ReDi (Kouzelis et al., 2025)	20.2	–
	JERI (Ours)	**9.7** (10.5)	**87.0**
DiT-XL	DiT (Peebles & Xie, 2023)	19.5	–
	REPA (Yu et al., 2025)	12.3	–
	ReDi (Kouzelis et al., 2025)	5.9	–
	JERI (Ours)	**3.9** (2.0)	**141.5**

representation that is still mostly noisy. As indicated by our theoretical analysis in Sec. 4.4, our design exhibits less divergence than ReDi when conditioned on clean representations.

6.2.2 Class-Unconditional Generation

We further evaluate unconditional generation. Unconditional generation is harder than class-conditional generation and provides a cleaner setup, avoiding the shortcut from labels to representations. According to Table 2, our JERI outperforms the baseline RCG significantly under the same compute. More importantly, our method also shows better representation quality, as indicated by a much lower representation FID (R-FID). This is consistent with our motivation and theoretical analysis: by partial joint coupling, JERI avoids the full imageless cost of RCG, where the representation can receive image contexts and achieve better consistency. The FID and IS gains persist at 1M iterations under the same training and evaluation setup.

Table 2: Unconditional ImageNet-1k generation with DiT-B. RCG and JERI use 400k iterations unless marked 1M. Green: gains over RCG at equal steps. *: ReDi (Kouzelis et al., 2025).

Model	FID ↓	IS ↑	R-FID ↓
DiT* (Peebles & Xie, 2023)	69.3	-	-
ReDi* (Kouzelis et al., 2025)	47.3	-	-
RCG (Li et al., 2024a)	13.1	57.3	16.5
JERI (Ours)	**12.1** (1.0)	**61.4** (4.1)	**13.3** (3.2)
RCG (1M)	9.2	75.4	–
JERI (1M)	**8.6** (0.6)	**79.2** (3.8)	–

6.3 Ablation Studies

We ablate the most important design choices of JERI on our ImageNet subset unconditional generation, as shown in Table 3. We highlight two conclusions. (1) Our *partial joint coupling* driven by theory is better than both extremes of $s = 0$ and $s = T$, with noticeable improvements; and (2) *teacher forcing* (Sec. 5.3) indeed stabilizes the training.

Table 3: Ablation study for JERI's design choices. Red numbers show the degradation relative to our final model. Both of our *partially joint schedule* and *teacher forcing* are crucial to better generation.

Model	FID ↓	IS ↑	R-FID ↓
JERI (Ours)	**25.0**	**45.9**	**22.3**
RCG ($s = 0$)	28.4 (3.4)	42.9 (3.0)	51.7 (29.4)
ReDi ($s = T$)	32.0 (7.0)	39.2 (6.7)	40.0 (17.7)
w/o Teacher Forcing	27.5 (2.5)	45.3 (0.6)	22.2

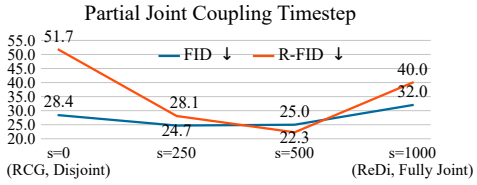


Figure 6: Generation quality across coupling timesteps with $T = 1000$. Both tested interior schedules improve image FID and representation FID over the two endpoint schedules, consistent with the measured trajectory cost in Fig. 4.

6.4 Actual Image Generation Performance Aligns with Our Trajectory Cost

We investigate whether image generation quality follows the trajectory cost $\mathcal{C}_{\text{JERI}}(s)$ in Fig. 4. Following Sec. 6.3, we analyze how the coupling timestep s affects generation quality. Suppose the total timestep in diffusion models $T = 1000$. Then $s = 0$ (RCG) and $s = 1000$ (ReDi) correspond to the two extremes of the coupling family, while $s = 250$ and $s = 500$ are interior couplings.

Delving into the details of Fig. 6, we find that both image and representation benefit from interior coupling. Compared with $s = 0$ (RCG), the evolution of the representation can better tailor to the image content and lead to higher quality at both ends, avoiding the imageless cost. Compared with $s = 1000$ (ReDi), our solution decouples the early noisy representation steps and avoids the noisy representation cost. For the two interior schedules, both $s = 250$ and $s = 500$ outperform the endpoint schedules in image and representation quality. We adopt $s = 500$ as a simple midpoint choice: it lies in the robust low-cost range identified by the probing analysis and avoids tuning the coupling timestep separately for each representation space. This trend is consistent with the trajectory-cost measurements in Fig. 4, where a broad interior range around $300 \leq s \leq 500$ remains preferable to both endpoints. Overall, the generation sweep supports the trajectory analysis: coupling should begin inside the denoising process.

6.5 Analysis

6.5.1 Improving with Data Diversity

Our JERI introduces a more general joint representation-image formulation for optimization, compared with the previous work (Li et al., 2024a). As it relies on less inductive bias, we naturally question *whether our JERI can better learn from harder and more diverse data*?

Table 4: Comparison of RCG and our JERI across different data diversities. Δ indicates the absolute improvement over RCG; larger is better. JERI has a larger advantage in representation quality, which also translates to better image quality in the more diverse settings.

Model	Diversity		
	First	Random	Furthest
RCG (Li et al., 2024a)	13.8	28.4	31.8
JERI (Ours)	13.8	**25.0**	**25.9**
Δ ↑	0.0	3.4	5.9

(a) Image FID (↓).

Model	Diversity		
	First	Random	Furthest
RCG (Li et al., 2024a)	13.1	51.7	44.1
JERI (Ours)	**12.2**	**22.3**	**22.6**
Δ ↑	0.9	29.4	21.5

(b) Representation FID (↓).

Curating Sets with Different Diversity. In addition to the random 100-class subset used in Sec. 6.2, which approximates a uniform distribution of classes, we prepare two additional settings that represent *less* and *more* diverse distributions. **(1)** We sample the first 100 classes from ImageNet for less diversity, which covers many semantically close objects, *e.g.*, multiple snake species. **(2)** For larger diversity, we calculate the average MoCov3 features for each class, and select the group of classes with the furthest distances using the Gonzalez algorithm (Gonzalez, 1985).

Results and Analysis. As in Table 4, our JERI has a larger advantage on the more diverse cases, suggesting that our formulation is indeed more flexible. Delving deeper into Table 4, we find that the FID increases with larger diversity, indicating that the generation task is becoming more challenging. In such scenarios, our JERI shows a stronger ability to synthesize representations (Table 4b), which effectively translates into better images (Table 4a). This further supports our partial-joint design.

6.5.2 Generalization to More Representation Spaces

In addition to data diversity, we investigate whether JERI is effective across different representation spaces. We compare JERI with RCG on three representation spaces: MoCov3 (Chen et al., 2021), DINOv2 (Oquab et al., 2023), and MAE (He et al., 2022). For all representation spaces, we ensure a consistent setup by applying PCA compression to a unified 256-dimensional feature, following the practice of ReDi (Kouzelis et al., 2025). More details are in Sec. D.

Table 5: Analysis of JERI on *unconditional* generation across three representation spaces (MoCov3 (Chen et al., 2021), DINOv2 (Oquab et al., 2023), MAE (He et al., 2022)). JERI consistently improves RCG, supporting the effectiveness of our partial-joint design.

Model		FID ↓	IS ↑
MoCov3	RCG (Li et al., 2024a)	28.4	42.9
	JERI (Ours)	**25.0** (3.4)	**45.9** (3.0)
DINOv2	RCG (Li et al., 2024a)	39.7	37.8
	JERI (Ours)	**23.3** (16.4)	**47.4** (9.6)
MAE	RCG (Li et al., 2024a)	54.4	26.4
	JERI (Ours)	**43.8** (10.6)	**30.0** (3.6)

As shown in Table 5, JERI consistently outperforms RCG across all three settings. Notably, the improvement is greater in more challenging spaces (DINOv2: −16.4 FID; MAE: −10.6 FID) than in the simpler MoCov3 (−3.4 FID), suggesting that our improvement to the partial-joint schedule is more beneficial as the representation space becomes harder to synthesize. This pattern aligns with our finding in Sec. 6.5.1, supporting the generality of our method.

6.6 Visualizations

Fig. 7 shows class-unconditional samples from our DiT-B. Despite the limited model capacity and training iterations, our JERI can capture the data distribution and generate valid images.



Figure 7: JERI's class-unconditional generation samples.

7 Conclusion

In this work, we study when representations and images should be coupled in representation-guided diffusion. We unify existing designs under a coupling-timestep view and show that both endpoints incur complementary information costs: denoising representations without image context and denoising images with noisy representations. This analysis motivates an intermediate coupling timestep, realized by JERI, which first advances representations and then jointly evolves representations and images. Experiments across conditional and unconditional generation, model scales, and representation spaces validate the effectiveness of our design.

AI USE STATEMENT

In this work, we used generative AI tools only for grammar checking and language polishing to improve the readability of the manuscript. We did not use generative AI tools for any tasks requiring disclosure under the ICLR policy. The research ideas, methodology, experiments, analysis, related work, and overall structure of the paper were developed and written by the authors. All AI-assisted edits were reviewed by the authors to ensure correctness and preservation of the intended meaning. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2018. URL https://arxiv.org/abs/1610.01644. 6

Alan Baade, Eric Ryan Chan, Kyle Sargent, Changan Chen, Justin Johnson, Ehsan Adeli, and Fei-Fei Li. Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. *arXiv preprint arXiv:2602.11401*, 2026. 1, 3

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *International Conference on Machine Learning*, 2025a. 3

Hao Chen, Ze Wang, Xiang Li, Ximeng Sun, Fangyi Chen, Jiang Liu, Jindong Wang, Bhiksha Raj, Zicheng Liu, and Emad Barsoum. Softvq-vae: Efficient 1-dimensional continuous tokenizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. 3

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *International Conference on Learning Representations*, 2025c. 3

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 1, 3, 9, 27, 29

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6, 7, 28

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3, 7

Weitao Du. Frequency-forcing: From scaling-as-time to soft frequency guidance. *arXiv preprint arXiv:2604.20902*, 2026. 3

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 3

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 1985. 9, 29

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 9

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. In *NeurIPS*, 2017. 7

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 28

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 7, 24, 27, 28

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025. 3

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025. 3

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 3, 24

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *CVPR*, 2024. 3

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021. 3

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7, 28

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3

Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. In *NeurIPS*, 2025. 1, 3, 4, 7, 8, 9, 27, 28

Samuel Lavoie, Michael Noukhovitch, and Aaron Courville. Compositional discrete latent code for high fidelity, productive diffusion models. *arXiv preprint arXiv:2507.12318*, 2025. 3, 27

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. 3, 27

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025. 3

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In *NeurIPS*, 2024a. 1, 3, 6, 7, 8, 9, 27, 28

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*, 2024b. 3

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Jindong Wang, Zhe Lin, and Bhiksha Raj. Xq-gan: An open-source image tokenization framework for autoregressive generation. *arXiv preprint arXiv:2412.01762*, 2024c. 3

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. 24

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7, 28

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 3, 9

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 3, 7, 8, 27, 28, 30

Kai Qiu, Xiang Li, Jason Kuen, Hao Chen, Xiaohao Xu, Jiuxiang Gu, Yinyi Luo, Bhiksha Raj, Zhe Lin, and Marios Savvides. Robust latent matters: Boosting image generation with sampling error synthesis. *arXiv preprint arXiv:2503.08354*, 2025. 3

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 7

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*, 2025. 3

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3, 24

timm. mini-imagenet. URL https://huggingface.co/datasets/timm/mini-imagenet. 7, 29

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026. 3

Chenyu Wang, Cai Zhou, Sharut Gupta, Zongyu Lin, Stefanie Jegelka, Stephen Bates, and Tommi Jaakkola. Learning diffusion models with flexible representation guidance. *arXiv preprint arXiv:2507.08980*, 2025a. 3

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025. 3

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025b. 3, 7, 27

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, et al. REPA works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *arXiv preprint arXiv:2505.16792*, 2025c. 3

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a. 3

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. In *NeurIPS*, 2025b. 1, 3, 4

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

Hu Yu, Hang Xu, Jie Huang, Zeyue Xue, Haoyang Huang, Nan Duan, and Feng Zhao. Rae-ar: Taming autoregressive models with representation autoencoders. *arXiv preprint arXiv:2604.01545*, 2026. 3

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025. 1, 3, 7, 26, 27

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 3

A PROOFS

A.1 ASSUMPTIONS

Throughout the appendix, all probabilities, conditional distributions, and expectations are taken under the true joint distribution induced by the clean pair (x_0, z_0) and the forward diffusion process. We write integrals for continuous variables; the same arguments apply to discrete variables by replacing integrals with sums.

Assumption A.1 (Forward process)**.** *For each timestep $t \in \{1, \dots, T\}$, the image $x_t \in \mathbb{R}^d$ and representation $z_t \in \mathbb{R}^k$ evolve according to independent variance-preserving forward processes:*

$$x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \xi_t^x, \tag{A.1}$$

$$z_t = \sqrt{\alpha_t}\, z_{t-1} + \sqrt{1-\alpha_t}\, \xi_t^z, \tag{A.2}$$

where $\alpha_t = 1 - \beta_t$, $\xi_t^x \sim \mathcal{N}(0, I_d)$, and $\xi_t^z \sim \mathcal{N}(0, I_k)$. The forward noises are independent across timesteps, independent across the image and representation branches, and independent of the clean pair (x_0, z_0).

Equivalently, for each fixed t,

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon_t^x, \tag{A.3}$$

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon_t^z, \tag{A.4}$$

where

$$\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i, \qquad \bar{\alpha}_0 := 1,$$

and $\varepsilon_t^x \sim \mathcal{N}(0, I_d)$, $\varepsilon_t^z \sim \mathcal{N}(0, I_k)$ are independent of (x_0, z_0) and of each other.

We consider noisy diffusion timesteps satisfying

$$0 < \alpha_t < 1, \qquad \bar{\alpha}_t > 0.$$

Assumption A.2 (Correlated images and representations)**.** *The clean image and representation have finite second moments and are correlated:*

$$\mathrm{Cov}(x_0, z_0) \neq \mathbf{0}.$$

Assumption A.3 (Gaussian specialization for closed forms)**.** *Only for the closed-form calculations, we additionally assume*

$$(x_0, z_0) \sim \mathcal{N}(0, \Sigma), \qquad \mathrm{Cov}(x_0) = I_d, \qquad \mathrm{Cov}(z_0) = I_k.$$

In the scalar case $d = k = 1$, we write

$$\rho := \mathrm{Cov}(x_0, z_0) \in (-1, 1) \setminus \{0\}.$$

Notation. The per-step information costs are

$$I_t^z := I(z_{t-1}; x_t \mid z_t), \qquad I_t^x := I(x_{t-1}; z_0 \mid x_t, z_t). \tag{A.5}$$

A.2 A GENERIC PYTHAGOREAN IDENTITY

We first prove a generic identity that will be applied as a fundamental tool to both image and representation denoising, used in later proofs, and it defines the cost of actual denoising and ideal denoising. Let U denote the variable to be denoised, V the information available to a restricted model, and Y the additional oracle information unavailable to the restricted model. The oracle transition is $p(u \mid y, v)$, while a restricted transition has the form $q(u \mid v)$.

Define the expected KL loss of q against the oracle transition as

$$\mathcal{L}(q) := \mathbb{E}_{(Y,V)\sim p(y,v)} \left[\mathrm{KL}\big(p(U \mid Y, V) \,\|\, q(U \mid V)\big) \right]. \tag{A.6}$$

Expanding the KL divergence, we obtain

$$\begin{aligned}\mathcal{L}(q) &= \int p(y,v) \int p(u \mid y,v) \log \frac{p(u \mid y,v)}{q(u \mid v)}\, du\, dy\, dv \\ &= \mathbb{E}_{(U,Y,V)\sim p(u,y,v)} \left[\log \frac{p(U \mid Y,V)}{q(U \mid V)}\right]. \end{aligned} \tag{A.7}$$

By the law of total probability, for every fixed v,

$$\begin{aligned}\mathbb{E}_{Y|V=v}\big[p(u \mid Y, V=v)\big] &= \int p(u \mid y,v) p(y \mid v)\, dy \\ &= \int p(u,y \mid v)\, dy \\ &= p(u \mid v). \end{aligned} \tag{A.8}$$

Thus averaging the oracle transition over the unavailable information Y gives exactly the Bayes-optimal restricted transition $p(u \mid v)$.

Now insert $p(U \mid V)$ into Eq. (A.7):

$$\mathcal{L}(q) = \mathbb{E}_{U,Y,V}\left[\log \frac{p(U \mid Y,V)}{p(U \mid V)}\right] + \mathbb{E}_{U,Y,V}\left[\log \frac{p(U \mid V)}{q(U \mid V)}\right]. \tag{A.9}$$

The first term is the conditional mutual information:

$$\begin{aligned} I(U;Y \mid V) &= \mathbb{E}_{U,Y,V}\left[\log \frac{p(U,Y \mid V)}{p(U \mid V)p(Y \mid V)}\right] \\ &= \mathbb{E}_{U,Y,V}\left[\log \frac{p(U \mid Y,V)p(Y \mid V)}{p(U \mid V)p(Y \mid V)}\right] \\ &= \mathbb{E}_{U,Y,V}\left[\log \frac{p(U \mid Y,V)}{p(U \mid V)}\right]. \end{aligned} \tag{A.10}$$

For the second term in Eq. (A.9), the logarithm depends only on (U,V), not on Y. Therefore,

$$\begin{aligned}\mathbb{E}_{U,Y,V}\left[\log \frac{p(U \mid V)}{q(U \mid V)}\right] &= \mathbb{E}_{U,V}\left[\log \frac{p(U \mid V)}{q(U \mid V)}\right] \\ &= \mathbb{E}_{V}\left[\mathrm{KL}\big(p(U \mid V) \,\|\, q(U \mid V)\big)\right]. \end{aligned} \tag{A.11}$$

Combining Eqs. (A.9)–(A.11), we obtain

$$\mathcal{L}(q) = I(U;Y \mid V) + \mathbb{E}_{V}\left[\mathrm{KL}\big(p(U \mid V) \,\|\, q(U \mid V)\big)\right]. \tag{A.12}$$

Since the second term is nonnegative and vanishes if and only if $q(U \mid V) = p(U \mid V)$ almost surely, we have

$$\inf_{q(U|V)} \mathcal{L}(q) = I(U;Y \mid V). \tag{A.13}$$

This defines our costs in the upcoming sections.

A.3 Imageless Cost for the Representation Branch via Pythagorean Identity

We instantiate the generic identity in Sec. A.2 with

$$(U,Y,V) = (z_{t-1}, x_t, z_t).$$

This compares the oracle transition $p(z_{t-1} \mid x_t, z_t)$, which has access to the image state x_t, with any restricted transition $q(z_{t-1} \mid z_t)$ that denoises the representation without image context.

Lemma A.1 (Representation-branch Pythagorean identity)**.** *Under Assumption A.1, for any conditional distribution $q(z_{t-1} \mid z_t)$,*

$$\begin{aligned}&\mathbb{E}_{(x_t,z_t)\sim p(x_t,z_t)}\left[\mathrm{KL}\big(p(z_{t-1} \mid x_t, z_t) \,\|\, q(z_{t-1} \mid z_t)\big)\right] \\ &\quad = I(z_{t-1}; x_t \mid z_t) + \mathbb{E}_{z_t\sim p(z_t)}\left[\mathrm{KL}\big(p(z_{t-1} \mid z_t) \,\|\, q(z_{t-1} \mid z_t)\big)\right]. \end{aligned} \tag{A.14}$$

Consequently,

$$\inf_{q(z_{t-1}|z_t)} \mathbb{E}_{(x_t,z_t)\sim p(x_t,z_t)} \left[\mathrm{KL}\big(p(z_{t-1} \mid x_t, z_t) \,\|\, q(z_{t-1} \mid z_t)\big)\right] = I(z_{t-1}; x_t \mid z_t) = I_t^z, \tag{A.15}$$

with the infimum attained at

$$q^\star(z_{t-1} \mid z_t) = p(z_{t-1} \mid z_t).$$

Proof. Equation (A.14) is Eq. (A.12) under the substitution displayed above. The second term in Eq. (A.14) is nonnegative and vanishes if and only if

$$q(z_{t-1} \mid z_t) = p(z_{t-1} \mid z_t)$$

almost surely. Therefore the minimum over all transitions that only condition on z_t is exactly

$$I(z_{t-1}; x_t \mid z_t) = I_t^z.$$

□

A.4 Noisy Representation Cost for the Image Branch via Pythagorean Identity

We instantiate the generic identity in Sec. A.2 with

$$(U, Y, V) = (x_{t-1}, z_0, (x_t, z_t)).$$

This compares the oracle transition $p(x_{t-1} \mid x_t, z_t, z_0)$, which has access to the clean representation z_0, with any restricted transition $q(x_{t-1} \mid x_t, z_t)$ that denoises the image using only the noisy representation state z_t.

Lemma A.2 (Image-branch Pythagorean identity)**.** *Under Assumption A.1, for any conditional distribution $q(x_{t-1} \mid x_t, z_t)$,*

$$\mathbb{E}_{(x_t,z_t,z_0)\sim p(x_t,z_t,z_0)} \left[\mathrm{KL}\big(p(x_{t-1} \mid x_t, z_t, z_0) \,\|\, q(x_{t-1} \mid x_t, z_t)\big)\right] = I(x_{t-1}; z_0 \mid x_t, z_t) + \mathbb{E}_{(x_t,z_t)\sim p(x_t,z_t)} \left[\mathrm{KL}\big(p(x_{t-1} \mid x_t, z_t) \,\|\, q(x_{t-1} \mid x_t, z_t)\big)\right]. \tag{A.16}$$

Consequently,

$$\inf_{q(x_{t-1}|x_t,z_t)} \mathbb{E}_{(x_t,z_t,z_0)\sim p(x_t,z_t,z_0)} \left[\mathrm{KL}\big(p(x_{t-1} \mid x_t, z_t, z_0) \,\|\, q(x_{t-1} \mid x_t, z_t)\big)\right] = I(x_{t-1}; z_0 \mid x_t, z_t) = I_t^x, \tag{A.17}$$

with the infimum attained at

$$q^\star(x_{t-1} \mid x_t, z_t) = p(x_{t-1} \mid x_t, z_t).$$

Proof. Equation (A.16) is Eq. (A.12) under the substitution displayed above. The second term in Eq. (A.16) is nonnegative and vanishes if and only if

$$q(x_{t-1} \mid x_t, z_t) = p(x_{t-1} \mid x_t, z_t)$$

almost surely. Therefore the minimum over all transitions that only condition on (x_t, z_t) is exactly

$$I(x_{t-1}; z_0 \mid x_t, z_t) = I_t^x.$$

□

A.5 Strict Positivity from Gaussian Forward Noising

The Pythagorean identities above reduce the endpoint gaps to conditional mutual informations. To show that these quantities are strictly positive, we use one elementary fact about Gaussian forward noising: observing a noisy version of a variable cannot completely screen off its genuine dependence with another variable. We state this fact once and then apply it to the two diffusion branches.

Lemma A.3 (Noisy forward steps do not screen off dependence)**.** *Let U and Y be random variables with arbitrary joint distribution. Let*

$$V = aU + \sigma\xi, \qquad a \neq 0, \quad \sigma > 0, \qquad \xi \sim \mathcal{N}(0, I), \tag{A.18}$$

where ξ is independent of (U, Y). If Y and U are not independent, then

$$I(U; Y \mid V) > 0. \tag{A.19}$$

Equivalently, if $Y \perp U \mid V$, then $Y \perp U$.

Proof. We prove the equivalent reverse implication. Since V is obtained from U by adding independent Gaussian noise, we have

$$Y \; - \; U \; - \; V, \tag{A.20}$$

and therefore

$$p(y \mid u, v) = p(y \mid u). \tag{A.21}$$

Assume $Y \perp U \mid V$. Then

$$p(y \mid u, v) = p(y \mid v). \tag{A.22}$$

Combining Eqs. (A.21) and (A.22), we get

$$p(y \mid u) = p(y \mid v) \tag{A.23}$$

for almost every (u, v) under the joint distribution of (U, V).

Because $\sigma > 0$, the conditional density $p(v \mid u)$ induced by Eq. (A.18) is Gaussian and strictly positive for every pair (u, v). Hence any two values u and u' can be compared through the same noisy value v. Using Eq. (A.23) twice gives

$$p(y \mid u) = p(y \mid v) = p(y \mid u') \tag{A.24}$$

for almost every u and u'. Thus $p(y \mid u)$ does not depend on u. Denote this common value by $r(y)$. Marginalizing over U gives

$$p(y) = \int p(y \mid u)p(u)\, du = \int r(y)p(u)\, du = r(y). \tag{A.25}$$

Therefore $p(y \mid u) = p(y)$, so $Y \perp U$.

Taking the contrapositive, if Y and U are not independent, then Y and U cannot be conditionally independent given V. Since conditional mutual information is zero if and only if conditional independence holds, this gives Eq. (A.19). □

Lemma A.4 (Noisy forward steps with side information)**.** *Let W be an additional conditioning variable. Suppose that, conditional on $W = w$,*

$$V = aU + \sigma\xi, \qquad a \neq 0, \quad \sigma > 0, \tag{A.26}$$

where $\xi \sim \mathcal{N}(0, I)$ is independent of (U, Y) given $W = w$. If

$$I(U; Y \mid W) > 0, \tag{A.27}$$

then

$$I(U; Y \mid V, W) > 0. \tag{A.28}$$

Equivalently, if $Y \perp U \mid (V, W)$, then $Y \perp U \mid W$.

Proof. We again prove the equivalent reverse implication. Fix a value w for which the conditional distribution given $W = w$ is well-defined. Conditional on $W = w$, the variables (U, Y, V) satisfy Lemma A.3. Hence

$$Y \perp U \mid V, W = w \quad \Longrightarrow \quad Y \perp U \mid W = w. \tag{A.29}$$

Since Eq. (A.29) holds for almost every w, we obtain

$$Y \perp U \mid (V, W) \quad \Longrightarrow \quad Y \perp U \mid W. \tag{A.30}$$

Taking the contrapositive of Eq. (A.30), if Y and U are not conditionally independent given W, then they cannot be conditionally independent after additionally conditioning on V. Equivalently,

$$I(U; Y \mid W) > 0 \quad \Longrightarrow \quad I(U; Y \mid V, W) > 0,$$

which proves Eq. (A.28). □

For Proposition 4.1, we use Lemma A.3 with $(U, Y, V) = (z_{t-1}, x_t, z_t)$, since z_t is obtained from z_{t-1} by one Gaussian forward noising step. For Proposition 4.2, we use Lemma A.4 to remove the noisy image step x_t while conditioning on z_t, and then use Lemma A.3 to remove the noisy representation step z_t.

A.6 Positivity of Imageless Cost and Noisy Representation Cost

Proposition (Proposition 4.1, restated)**.** *Under Assumptions A.1 and A.2, for every noisy diffusion timestep t with $0 < \alpha_t < 1$ and $0 < \bar{\alpha}_t < 1$,*

$$I_t^z = I(z_{t-1}; x_t \mid z_t) > 0.$$

Equivalently,

$$p(z_{t-1} \mid x_t, z_t) \neq p(z_{t-1} \mid z_t)$$

on a set of positive probability.

Proof. We argue by contradiction. Suppose

$$I(z_{t-1}; x_t \mid z_t) = 0. \tag{A.31}$$

Then $x_t \perp z_{t-1} \mid z_t$. On the other hand, the representation forward step is $z_t = \sqrt{\alpha_t}\, z_{t-1} + \sqrt{1-\alpha_t}\, \xi_t^z$, where ξ_t^z is independent of (z_{t-1}, x_t). Since $0 < \alpha_t < 1$, this is a noisy forward step from z_{t-1} to z_t. Applying Lemma A.3 to Eq. (A.31) gives

$$x_t \perp z_{t-1}. \tag{A.32}$$

It remains to show that Eq. (A.32) contradicts the clean correlation assumption. By the marginal form of the forward process, $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon_t^x, \quad z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, z_0 + \sqrt{1-\bar{\alpha}_{t-1}}\, \varepsilon_{t-1}^z$, where the noises are independent of (x_0, z_0) and of each other. Therefore,

$$\mathrm{Cov}(x_t, z_{t-1}) = \sqrt{\bar{\alpha}_t \bar{\alpha}_{t-1}}\, \mathrm{Cov}(x_0, z_0). \tag{A.33}$$

By Assumption A.2, $\mathrm{Cov}(x_0, z_0) \neq \mathbf{0}$, and by Assumption A.1, $\bar{\alpha}_t > 0$ and $\bar{\alpha}_{t-1} > 0$. Hence

$$\mathrm{Cov}(x_t, z_{t-1}) \neq \mathbf{0}.$$

Thus x_t and z_{t-1} cannot be independent, contradicting Eq. (A.32). Therefore,

$$I(z_{t-1}; x_t \mid z_t) > 0.$$

The equivalence to

$$p(z_{t-1} \mid x_t, z_t) \neq p(z_{t-1} \mid z_t)$$

on a set of positive probability follows from the standard equivalence between zero conditional mutual information and conditional independence. □

Proposition (Proposition 4.2, restated)**.** *Under Assumptions A.1 and A.2, for every noisy diffusion timestep t with $0 < \alpha_t < 1$ and $0 < \bar{\alpha}_t < 1$,*

$$I_t^x = I(x_{t-1}; z_0 \mid x_t, z_t) > 0.$$

Equivalently,

$$p(x_{t-1} \mid x_t, z_t, z_0) \neq p(x_{t-1} \mid x_t, z_t)$$

on a set of positive probability.

Proof. We again argue by contradiction. Suppose

$$I(x_{t-1}; z_0 \mid x_t, z_t) = 0. \tag{A.34}$$

Then $z_0 \perp x_{t-1} \mid (x_t, z_t)$. Since $x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \xi_t^x$ with ξ_t^x independent of (x_{t-1}, z_0, z_t), x_t is a noisy forward step from x_{t-1} even after conditioning on z_t. Lemma A.4 therefore removes this noisy image step from Eq. (A.34), yielding

$$z_0 \perp x_{t-1} \mid z_t. \tag{A.35}$$

Next, z_t is itself a noisy forward version of the clean representation: $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon_t^z$, where ε_t^z is independent of (x_{t-1}, z_0). Since $0 < \bar{\alpha}_t < 1$, Lemma A.3 applied to Eq. (A.35) gives

$$x_{t-1} \perp z_0. \tag{A.36}$$

This independence contradicts the clean correlation assumption. Indeed, by the marginal forward process, $x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, x_0 + \sqrt{1-\bar{\alpha}_{t-1}}\, \varepsilon_{t-1}^x$, with ε_{t-1}^x independent of (x_0, z_0). Hence

$$\mathrm{Cov}(x_{t-1}, z_0) = \sqrt{\bar{\alpha}_{t-1}}\, \mathrm{Cov}(x_0, z_0). \tag{A.37}$$

By Assumption A.2, $\mathrm{Cov}(x_0, z_0) \neq \mathbf{0}$, and by Assumption A.1, $\bar{\alpha}_{t-1} > 0$. Therefore

$$\mathrm{Cov}(x_{t-1}, z_0) \neq \mathbf{0},$$

so x_{t-1} and z_0 cannot be independent, contradicting Eq. (A.36). Thus,

$$I(x_{t-1}; z_0 \mid x_t, z_t) > 0.$$

The equivalence to

$$p(x_{t-1} \mid x_t, z_t, z_0) \neq p(x_{t-1} \mid x_t, z_t)$$

on a set of positive probability again follows from the equivalence between zero conditional mutual information and conditional independence. □

A.7 Gaussian Closed Forms

We now give closed-form expressions for I_t^z and I_t^x under the Gaussian-data specialization in Assumption A.3. This is used for our probing in Sec. 4.4.

Proposition A.1 (Scalar Gaussian closed forms)**.** *Assume Assumptions A.1–A.3 with $d = k = 1$. Let*

$$\rho = \mathrm{Cov}(x_0, z_0) \in (-1, 1) \setminus \{0\}.$$

Then

$$I_t^z = \frac{1}{2} \log \frac{1 - \bar{\alpha}_t^2 \rho^2}{1 - \bar{\alpha}_t \bar{\alpha}_{t-1} \rho^2}, \tag{A.38}$$

and

$$I_t^x = \frac{1}{2} \log \frac{(1 - \bar{\alpha}_t \bar{\alpha}_{t-1} \rho^2)(1 - \bar{\alpha}_t \rho^2)}{(1 - \bar{\alpha}_{t-1} \rho^2)(1 - \bar{\alpha}_t^2 \rho^2)}. \tag{A.39}$$

Both quantities are strictly positive whenever $\rho \neq 0$ and the forward step is non-degenerate.

Proof. We first compute I_t^z. Under Assumption A.3, the triple (z_{t-1}, x_t, z_t) is jointly Gaussian. Its covariance matrix is

$$\Sigma_{(z_{t-1}, x_t, z_t)} = \begin{pmatrix} 1 & \sqrt{\bar{\alpha}_{t-1}\bar{\alpha}_t}\,\rho & \sqrt{\alpha_t} \\ \sqrt{\bar{\alpha}_{t-1}\bar{\alpha}_t}\,\rho & 1 & \bar{\alpha}_t \rho \\ \sqrt{\alpha_t} & \bar{\alpha}_t \rho & 1 \end{pmatrix}. \tag{A.40}$$

Conditioning (z_{t-1}, x_t) on z_t gives

$$\mathrm{Var}(z_{t-1} \mid z_t) = 1 - \alpha_t, \tag{A.41}$$

$$\mathrm{Var}(x_t \mid z_t) = 1 - \bar{\alpha}_t^2 \rho^2, \tag{A.42}$$

$$\begin{aligned} \mathrm{Cov}(z_{t-1}, x_t \mid z_t) &= \sqrt{\bar{\alpha}_{t-1}\bar{\alpha}_t}\,\rho - \sqrt{\alpha_t}\,\bar{\alpha}_t \rho \\ &= \sqrt{\alpha_t}\bar{\alpha}_{t-1}\rho - \sqrt{\alpha_t}\alpha_t\bar{\alpha}_{t-1}\rho \\ &= \sqrt{\alpha_t}\bar{\alpha}_{t-1}(1-\alpha_t)\rho. \end{aligned} \tag{A.43}$$

For a bivariate Gaussian pair (A, B) conditioned on C,

$$I(A; B \mid C) = \frac{1}{2} \log \frac{\mathrm{Var}(A \mid C)\, \mathrm{Var}(B \mid C)}{\det \Sigma_{(A,B)|C}}.$$

Therefore,

$$I_t^z = \frac{1}{2} \log \frac{(1-\alpha_t)(1-\bar{\alpha}_t^2\rho^2)}{(1-\alpha_t)(1-\bar{\alpha}_t^2\rho^2) - \alpha_t \bar{\alpha}_{t-1}^2 (1-\alpha_t)^2 \rho^2}. \tag{A.44}$$

Canceling the common factor $(1-\alpha_t)$ in the denominator,

$$I_t^z = \frac{1}{2}\log\frac{1-\bar{\alpha}_t^2\rho^2}{1-\bar{\alpha}_t^2\rho^2-\alpha_t\bar{\alpha}_{t-1}^2(1-\alpha_t)\rho^2}. \tag{A.45}$$

Using $\bar{\alpha}_t = \alpha_t\bar{\alpha}_{t-1}$,

$$\begin{aligned}\bar{\alpha}_t^2 + \alpha_t\bar{\alpha}_{t-1}^2(1-\alpha_t) &= \alpha_t^2\bar{\alpha}_{t-1}^2 + \alpha_t\bar{\alpha}_{t-1}^2(1-\alpha_t)\\ &= \alpha_t\bar{\alpha}_{t-1}^2 = \bar{\alpha}_t\bar{\alpha}_{t-1}.\end{aligned} \tag{A.46}$$

Hence

$$I_t^z = \frac{1}{2}\log\frac{1-\bar{\alpha}_t^2\rho^2}{1-\bar{\alpha}_t\bar{\alpha}_{t-1}\rho^2}.$$

We next compute I_t^x. Under Assumption A.3, the quartet (x_{t-1}, z_0, x_t, z_t) is jointly Gaussian. Partition

$$A = (x_{t-1}, z_0), \qquad C = (x_t, z_t).$$

The relevant covariance blocks are

$$\Sigma_{AA} = \begin{pmatrix} 1 & \sqrt{\bar{\alpha}_{t-1}}\rho \\ \sqrt{\bar{\alpha}_{t-1}}\rho & 1 \end{pmatrix}, \tag{A.47}$$

$$\Sigma_{CC} = \begin{pmatrix} 1 & \bar{\alpha}_t\rho \\ \bar{\alpha}_t\rho & 1 \end{pmatrix}, \tag{A.48}$$

$$\Sigma_{AC} = \begin{pmatrix} \sqrt{\alpha_t} & \sqrt{\bar{\alpha}_{t-1}\bar{\alpha}_t}\rho \\ \sqrt{\bar{\alpha}_t}\rho & \sqrt{\bar{\alpha}_t} \end{pmatrix}. \tag{A.49}$$

Let

$$D := 1-\bar{\alpha}_t^2\rho^2.$$

Then

$$\Sigma_{CC}^{-1} = \frac{1}{D}\begin{pmatrix} 1 & -\bar{\alpha}_t\rho \\ -\bar{\alpha}_t\rho & 1 \end{pmatrix}.$$

The conditional covariance is

$$\Sigma_{A|C} = \Sigma_{AA} - \Sigma_{AC}\Sigma_{CC}^{-1}\Sigma_{CA}.$$

A direct multiplication gives

$$[\Sigma_{A|C}]_{11} = \frac{(1-\alpha_t)(1-\bar{\alpha}_t\bar{\alpha}_{t-1}\rho^2)}{D}, \tag{A.50}$$

$$[\Sigma_{A|C}]_{22} = \frac{(1-\bar{\alpha}_t)(1-\bar{\alpha}_t\rho^2)}{D}, \tag{A.51}$$

$$[\Sigma_{A|C}]_{12} = \frac{\sqrt{\bar{\alpha}_{t-1}}\rho(1-\alpha_t)(1-\bar{\alpha}_t)}{D}. \tag{A.52}$$

Thus,

$$I_t^x = \frac{1}{2}\log\frac{[\Sigma_{A|C}]_{11}[\Sigma_{A|C}]_{22}}{\det\Sigma_{A|C}}. \tag{A.53}$$

Substituting the entries above and simplifying the determinant gives

$$I_t^x = \frac{1}{2}\log\frac{(1-\bar{\alpha}_t\bar{\alpha}_{t-1}\rho^2)(1-\bar{\alpha}_t\rho^2)}{(1-\bar{\alpha}_{t-1}\rho^2)(1-\bar{\alpha}_t^2\rho^2)}. \tag{A.54}$$

Finally, strict positivity follows by comparing numerator and denominator. For I_t^z, since $0<\alpha_t<1$ implies $\bar{\alpha}_t < \bar{\alpha}_{t-1}$,

$$\bar{\alpha}_t^2 < \bar{\alpha}_t\bar{\alpha}_{t-1}.$$

Thus

$$1-\bar{\alpha}_t^2\rho^2 > 1-\bar{\alpha}_t\bar{\alpha}_{t-1}\rho^2,$$

so $I_t^z > 0$.

For I_t^x, the numerator minus the denominator inside the logarithm is

$$\begin{aligned}&(1-\bar{\alpha}_t\bar{\alpha}_{t-1}\rho^2)(1-\bar{\alpha}_t\rho^2) - (1-\bar{\alpha}_{t-1}\rho^2)(1-\bar{\alpha}_t^2\rho^2)\\ &\quad= \bar{\alpha}_{t-1}\rho^2(1-\alpha_t)(1-\bar{\alpha}_t),\end{aligned} \tag{A.55}$$

which is strictly positive for $\rho\neq 0$, $0<\alpha_t<1$, and $\bar{\alpha}_t<1$. Therefore $I_t^x > 0$. □

A.8 Trajectory Information Costs

For $1 \le t \le T$ and $0 \le \tau \le T$, define the two-time information costs

$$I^x_{t,\tau} := I(x_{t-1}; z_0 \mid x_t, z_\tau), \qquad I^z_{t,\tau} := I(z_{t-1}; x_\tau \mid z_t). \tag{A.56}$$

At $\tau = t$, these recover the per-step costs in Eq. (A.5): $I^x_{t,t} = I^x_t$ and $I^z_{t,t} = I^z_t$. Moreover, $I^x_{t,0} = I(x_{t-1}; z_0 \mid x_t, z_0) = 0$.

For a coupling timestep $s \in \{0, \dots, T\}$, define the representation timestep paired with image timestep t by

$$\tau_s(t) := \left\lfloor \frac{s}{T} t \right\rfloor. \tag{A.57}$$

The two-time JERI cost is

$$\mathcal{C}_{\mathrm{JERI}}(s) := \sum_{u=s+1}^{T} I^z_u + \sum_{t=1}^{T} I^x_{t,\tau_s(t)}. \tag{A.58}$$

The first sum is the representation-alone cost before coupling. The second sum is the image-branch cost from conditioning on the paired representation state $z_{\tau_s(t)}$ instead of clean z_0.

At $s = 0$, $\tau_0(t) = 0$, so the image-branch sum vanishes and $\mathcal{C}_{\mathrm{JERI}}(0) = \sum_{t=1}^{T} I^z_t$. At $s = T$, $\tau_T(t) = t$, so the representation-alone sum is empty and $\mathcal{C}_{\mathrm{JERI}}(T) = \sum_{t=1}^{T} I^x_t$. These are the RCG and ReDi/REG endpoint costs, respectively.

For the proof, we use a one-time reduction that places the two costs on a common timestep axis:

$$\mathcal{C}(s) := \sum_{t=s+1}^{T} I^z_t + \sum_{t=1}^{s} I^x_t. \tag{A.59}$$

For every $s \in \{0, \dots, T\}$,

$$\begin{aligned}
\mathcal{C}_{\mathrm{JERI}}(s) - \mathcal{C}(s) &= \left[\sum_{u=s+1}^{T} I^z_u + \sum_{t=1}^{T} I^x_{t,\tau_s(t)} \right] - \left[\sum_{u=s+1}^{T} I^z_u + \sum_{t=1}^{s} I^x_t \right] \\
&= \sum_{t=1}^{T} I^x_{t,\tau_s(t)} - \sum_{t=1}^{s} I^x_t \\
&= \sum_{t=1}^{s} \left(I^x_{t,\tau_s(t)} - I^x_t \right) + \sum_{t=s+1}^{T} I^x_{t,\tau_s(t)} \\
&:= \Delta(s).
\end{aligned} \tag{A.60}$$

The one-time reduction shares the endpoint values of $\mathcal{C}_{\mathrm{JERI}}$ while aligning both costs on a common timestep axis. This alignment enables the monotonicity and telescoping proof of endpoint improvements. The bridge identity $\mathcal{C}_{\mathrm{JERI}}(s) = \mathcal{C}(s) + \Delta(s)$ then transfers the argument back to the actual two-time JERI schedule: after proving endpoint improvements for $\mathcal{C}(s)$, we only need to check that these improvements remain positive after the exact correction $\Delta(s)$.

A.9 Forward-Process Conditions for Interior Coupling

The two endpoints are excluded by two different comparisons. Moving from $s = 0$ to an interior timestep removes representation-alone denoising cost and introduces noisy-representation cost for the image branch. Moving from $s = T$ to an interior timestep removes noisy-representation cost and introduces representation-alone denoising cost. For $s \in \{1, \dots, T-1\}$, define

$$m_0(s) := 2I(x_s; z_0 \mid z_s) - I(x_0; z_0 \mid z_s), \tag{A.61}$$

$$m_T(s) := I(x_s; z_0 \mid z_s) - I(x_T; z_0 \mid z_s) - I(x_s; z_s \mid z_T). \tag{A.62}$$

The quantity $m_0(s)$ lower-bounds the improvement relative to the $s = 0$ endpoint: $I(x_s; z_0 \mid z_s)$ lower-bounds the representation-alone cost removed by coupling, while $I(x_0; z_0 \mid z_s) - I(x_s; z_0 \mid z_s)$

upper-bounds the noisy-representation cost introduced on the image branch. Similarly, $m_T(s)$ lower-bounds the improvement relative to the $s = T$ endpoint: $I(x_s; z_0 \mid z_s) - I(x_T; z_0 \mid z_s)$ lower-bounds the noisy-representation cost removed from the image branch, while $I(x_s; z_s \mid z_T)$ upper-bounds the representation-alone cost introduced before coupling. Conditions (H1) and (H2) require these two lower bounds to be positive.

Lemma A.5 (Two-time identities)**.** *Under Assumption A.1, whenever the displayed quantities are finite, for every $1 \le t \le T$ and every $0 \le a, \tau \le T$,*

$$I(x_a; z_0 \mid z_\tau) = I(x_a; z_0) - I(x_a; z_\tau), \tag{A.63}$$
$$I^z_{t,\tau} = I(z_{t-1}; x_\tau) - I(z_t; x_\tau), \tag{A.64}$$
$$I^x_{t,\tau} = I(x_{t-1}; z_0 \mid z_\tau) - I(x_t; z_0 \mid z_\tau). \tag{A.65}$$

Proof. For Eq. (A.63), the chain rule gives

$$\begin{aligned} I(x_a; z_0, z_\tau) &= I(x_a; z_\tau) + I(x_a; z_0 \mid z_\tau) \\ &= I(x_a; z_0) + I(x_a; z_\tau \mid z_0). \end{aligned} \tag{A.66}$$

Under Assumption A.1, z_τ is generated from z_0 using representation forward noise independent of the image branch, so $I(x_a; z_\tau \mid z_0) = 0$. Rearranging Eq. (A.66) proves Eq. (A.63).

For Eq. (A.64), the chain rule gives

$$\begin{aligned} I(z_{t-1}, z_t; x_\tau) &= I(z_t; x_\tau) + I(z_{t-1}; x_\tau \mid z_t) \\ &= I(z_{t-1}; x_\tau) + I(z_t; x_\tau \mid z_{t-1}). \end{aligned} \tag{A.67}$$

The transition from z_{t-1} to z_t uses representation forward noise independent of x_τ, so $I(z_t; x_\tau \mid z_{t-1}) = 0$. Since $I(z_{t-1}; x_\tau \mid z_t) = I^z_{t,\tau}$ by Eq. (A.56), rearranging Eq. (A.67) proves Eq. (A.64).

For Eq. (A.65), apply the chain rule conditionally on z_τ:

$$\begin{aligned} I(x_{t-1}, x_t; z_0 \mid z_\tau) &= I(x_t; z_0 \mid z_\tau) + I(x_{t-1}; z_0 \mid x_t, z_\tau) \\ &= I(x_{t-1}; z_0 \mid z_\tau) + I(x_t; z_0 \mid x_{t-1}, z_\tau). \end{aligned} \tag{A.68}$$

The transition from x_{t-1} to x_t uses image forward noise independent of (z_0, z_τ), so $I(x_t; z_0 \mid x_{t-1}, z_\tau) = 0$. Since $I(x_{t-1}; z_0 \mid x_t, z_\tau) = I^x_{t,\tau}$ by Eq. (A.56), rearranging Eq. (A.68) proves Eq. (A.65). □

Lemma A.6 (Monotonicity in the conditioning timestep)**.** *Under Assumption A.1, for every $1 \le t \le T$, the map $\tau \mapsto I^x_{t,\tau}$ is non-decreasing and the map $\tau \mapsto I^z_{t,\tau}$ is non-increasing on $\{0, \dots, T\}$.*

Proof. Let $0 \le \tau \le \tau' \le T$. By composing the forward transitions,

$$\begin{aligned} z_{\tau'} &= \sqrt{\frac{\bar\alpha_{\tau'}}{\bar\alpha_\tau}}\, z_\tau + \sqrt{1 - \frac{\bar\alpha_{\tau'}}{\bar\alpha_\tau}}\, \varepsilon^z_{\tau'|\tau}, \\ x_{\tau'} &= \sqrt{\frac{\bar\alpha_{\tau'}}{\bar\alpha_\tau}}\, x_\tau + \sqrt{1 - \frac{\bar\alpha_{\tau'}}{\bar\alpha_\tau}}\, \varepsilon^x_{\tau'|\tau}, \end{aligned} \tag{A.69}$$

where $\varepsilon^z_{\tau'|\tau} \sim \mathcal{N}(0, I_k)$ and $\varepsilon^x_{\tau'|\tau} \sim \mathcal{N}(0, I_d)$ are independent of the states at time τ, independent of the opposite branch, and independent of each other.

For the image-branch cost, expand $I(x_{t-1}; z_0, z_\tau \mid x_t)$ in two ways:

$$\begin{aligned} I(x_{t-1}; z_0, z_\tau \mid x_t) &= I(x_{t-1}; z_\tau \mid x_t) + I(x_{t-1}; z_0 \mid x_t, z_\tau) \\ &= I(x_{t-1}; z_0 \mid x_t) + I(x_{t-1}; z_\tau \mid x_t, z_0). \end{aligned} \tag{A.70}$$

Under Assumption A.1, $I(x_{t-1}; z_\tau \mid x_t, z_0) = 0$. Therefore

$$I^x_{t,\tau} = I(x_{t-1}; z_0 \mid x_t) - I(x_{t-1}; z_\tau \mid x_t). \tag{A.71}$$

Since $z_{\tau'}$ is obtained from z_τ through the independent forward transition in Eq. (A.69), the conditional data-processing inequality gives

$$I(x_{t-1}; z_{\tau'} \mid x_t) \le I(x_{t-1}; z_\tau \mid x_t). \tag{A.72}$$

Substituting this inequality into Eq. (A.71) gives $I^x_{t,\tau'} \geq I^x_{t,\tau}$.

For the representation-branch cost, $x_{\tau'}$ is obtained from x_τ through the independent forward transition in Eq. (A.69). The conditional data-processing inequality gives

$$I(z_{t-1}; x_{\tau'} \mid z_t) \leq I(z_{t-1}; x_\tau \mid z_t), \tag{A.73}$$

which is $I^z_{t,\tau'} \leq I^z_{t,\tau}$ by Eq. (A.56). □

Lemma A.7 (Endpoint lower bounds)**.** *Under Assumption A.1, whenever the displayed quantities are finite, for every $s \in \{1, \dots, T-1\}$,*

$$\mathcal{C}(0) - \mathcal{C}(s) \geq m_0(s), \tag{A.74}$$

and

$$\mathcal{C}(T) - \mathcal{C}(s) \geq m_T(s). \tag{A.75}$$

Proof. By Eq. (A.59),

$$\mathcal{C}(0) - \mathcal{C}(s) = \sum_{t=1}^{s} \left(I^z_t - I^x_t\right). \tag{A.76}$$

For $1 \leq t \leq s$, Lemma A.6 gives $I^z_t = I^z_{t,t} \geq I^z_{t,s}$ and $I^x_t = I^x_{t,t} \leq I^x_{t,s}$. Hence

$$\mathcal{C}(0) - \mathcal{C}(s) \geq \sum_{t=1}^{s} I^z_{t,s} - \sum_{t=1}^{s} I^x_{t,s}. \tag{A.77}$$

Using Eq. (A.64),

$$\begin{aligned}\sum_{t=1}^{s} I^z_{t,s} &= [I(z_0; x_s) - I(z_1; x_s)] + [I(z_1; x_s) - I(z_2; x_s)] + \cdots \\ &\quad + [I(z_{s-1}; x_s) - I(z_s; x_s)] \\ &= I(z_0; x_s) - I(z_s; x_s) \\ &= I(x_s; z_0 \mid z_s),\end{aligned} \tag{A.78}$$

where the final equality follows from Eq. (A.63). Using Eq. (A.65),

$$\begin{aligned}\sum_{t=1}^{s} I^x_{t,s} &= [I(x_0; z_0 \mid z_s) - I(x_1; z_0 \mid z_s)] + \cdots \\ &\quad + [I(x_{s-1}; z_0 \mid z_s) - I(x_s; z_0 \mid z_s)] \\ &= I(x_0; z_0 \mid z_s) - I(x_s; z_0 \mid z_s).\end{aligned} \tag{A.79}$$

Substituting Eqs. (A.78) and (A.79) into Eq. (A.77) gives

$$\begin{aligned}\mathcal{C}(0) - \mathcal{C}(s) &\geq 2I(x_s; z_0 \mid z_s) - I(x_0; z_0 \mid z_s) \\ &= m_0(s),\end{aligned} \tag{A.80}$$

which proves Eq. (A.74).

Similarly, Eq. (A.59) gives

$$\mathcal{C}(T) - \mathcal{C}(s) = \sum_{t=s+1}^{T} \left(I^x_t - I^z_t\right). \tag{A.81}$$

For $s+1 \leq t \leq T$, Lemma A.6 gives $I^x_t = I^x_{t,t} \geq I^x_{t,s}$ and $I^z_t = I^z_{t,t} \leq I^z_{t,s}$. Hence

$$\mathcal{C}(T) - \mathcal{C}(s) \geq \sum_{t=s+1}^{T} I^x_{t,s} - \sum_{t=s+1}^{T} I^z_{t,s}. \tag{A.82}$$

Using Eq. (A.65),

$$\begin{aligned}\sum_{t=s+1}^{T} I^x_{t,s} &= [I(x_s; z_0 \mid z_s) - I(x_{s+1}; z_0 \mid z_s)] + \cdots \\ &\quad + [I(x_{T-1}; z_0 \mid z_s) - I(x_T; z_0 \mid z_s)] \\ &= I(x_s; z_0 \mid z_s) - I(x_T; z_0 \mid z_s).\end{aligned} \tag{A.83}$$

Using Eq. (A.64),

$$\sum_{t=s+1}^{T} I^z_{t,s} = [I(z_s; x_s) - I(z_{s+1}; x_s)] + \cdots + [I(z_{T-1}; x_s) - I(z_T; x_s)] = I(z_s; x_s) - I(z_T; x_s). \tag{A.84}$$

The chain rule gives

$$I(x_s; z_s, z_T) = I(x_s; z_s) + I(x_s; z_T \mid z_s) = I(x_s; z_T) + I(x_s; z_s \mid z_T). \tag{A.85}$$

Under Assumption A.1, $I(x_s; z_T \mid z_s) = 0$, so

$$I(z_s; x_s) - I(z_T; x_s) = I(x_s; z_s \mid z_T). \tag{A.86}$$

Substituting Eqs. (A.83) and (A.86) into Eq. (A.82) gives

$$\mathcal{C}(T) - \mathcal{C}(s) \geq I(x_s; z_0 \mid z_s) - I(x_T; z_0 \mid z_s) - I(x_s; z_s \mid z_T) = m_T(s), \tag{A.87}$$

which proves Eq. (A.75). □

Proposition (Interior optimum for the one-time cost)**.** *Under Assumption A.1, suppose all mutual information terms below are finite. Suppose there exist* $s_0, s_T \in \{1, \ldots, T-1\}$ *such that*

$$m_0(s_0) > 0, \tag{H1}$$

and

$$m_T(s_T) > 0. \tag{H2}$$

Then both endpoints are strictly suboptimal for $\mathcal{C}(s)$*, and every global minimizer of* $\mathcal{C}(s)$ *lies in* $\{1, \ldots, T-1\}$.

Proof. By Lemma A.7,

$$\mathcal{C}(0) - \mathcal{C}(s_0) \geq m_0(s_0) > 0, \tag{A.88}$$

and

$$\mathcal{C}(T) - \mathcal{C}(s_T) \geq m_T(s_T) > 0. \tag{A.89}$$

Thus neither endpoint is a minimizer. Since $\mathcal{C}$ is defined on the finite set $\{0, 1, \ldots, T\}$, it attains a global minimum, and every global minimizer lies in $\{1, \ldots, T-1\}$. □

Theorem (Theorem 4.1, restated)**.** *Under Assumption A.1, suppose all mutual information terms in Eqs. (A.61), (A.62), and (A.60) are finite. If there exist* $s_0, s_T \in \{1, \ldots, T-1\}$ *such that*

$$\text{(C1) } m_0(s_0) - \Delta(s_0) > 0, \ \text{(C2) } m_T(s_T) - \Delta(s_T) > 0. \tag{A.90}$$

then both endpoints are strictly suboptimal for $\mathcal{C}_{\mathrm{JERI}}(s)$*, and every global minimizer of* $\mathcal{C}_{\mathrm{JERI}}(s)$ *lies in* $\{1, \ldots, T-1\}$.

Proof. By Eqs. (A.58), (A.59), and (A.60),

$$\mathcal{C}_{\mathrm{JERI}}(0) - \mathcal{C}_{\mathrm{JERI}}(s_0) = \mathcal{C}(0) - \mathcal{C}(s_0) - \Delta(s_0) \geq m_0(s_0) - \Delta(s_0) > 0, \tag{A.91}$$

where the inequality follows from Lemma A.7. Similarly,

$$\mathcal{C}_{\mathrm{JERI}}(T) - \mathcal{C}_{\mathrm{JERI}}(s_T) = \mathcal{C}(T) - \mathcal{C}(s_T) - \Delta(s_T) \geq m_T(s_T) - \Delta(s_T) > 0. \tag{A.92}$$

Thus neither endpoint is a minimizer. Since $\mathcal{C}_{\mathrm{JERI}}$ is defined on the finite set $\{0, 1, \ldots, T\}$, it attains a global minimum, and every global minimizer lies in $\{1, \ldots, T-1\}$. □

Conditions (C1) and (C2) have the same endpoint interpretation as the margins above, but for the actual two-time JERI schedule. Condition (C1) states that, after the warped-pairing correction $\Delta(s)$, leaving RCG still removes more Imageless Cost than it introduces Noisy Representation Cost. Condition (C2) states the analogous trade-off when leaving ReDi/REG. Since every term is a mutual information of the forward process, Sec. B measures these quantities directly on ImageNet directions selected from SSL representations.

A.10 Trajectory Cost and Image Generation

Trajectory cost depends only on the forward process, which is why it can be probed on ImageNet (Sec. 4.4, Appendix B). The final image KL also depends on the sampler and has a closed form only in the Gaussian setting of Assumption A.3. This section uses that setting to connect the two: we express the errors of finite-step samplers through the information terms of trajectory cost, and compute the final image KL of DDPM and Heun sampling for every coupling timestep.

Trajectory cost and sampling error. Seeing the other branch makes each reverse step more certain. For Gaussians, this gain is exactly an information term: conditional mutual information is half the log-ratio of covariance determinants before and after conditioning. Without the other branch, the reverse steps have covariances $\mathrm{Cov}(x_{t-1} \mid x_t) = \beta_t I_d$ and $\mathrm{Cov}(z_{u-1} \mid z_u) = \beta_u I_k$ under Assumption A.3. Since z_u arises from z_v, and z_v from z_0, by independent forward noise (Assumption A.1), the mutual-information chain rule and Eq. (A.56) give the gains

$$\begin{aligned} \log\det(\beta_t I_d) - \log\det \mathrm{Cov}(x_{t-1} \mid x_t, z_u) &= 2\, I(x_{t-1}; z_u \mid x_t) \\ &= 2\big[I(x_{t-1}; z_0 \mid x_t) - I^x_{t,u}\big], \qquad \text{(A.93)} \\ \log\det(\beta_u I_k) - \log\det \mathrm{Cov}(z_{u-1} \mid z_u, x_t) &= 2\, I(z_{u-1}; x_t \mid z_u) = 2\, I^z_{u,t}, \qquad \text{(A.94)} \\ I(x_{t-1}; z_v \mid x_t, z_u) &= I^x_{t,u} - I^x_{t,v}. \qquad \text{(A.95)} \end{aligned}$$

Both gains in Eqs. (A.93)–(A.94) shrink as the other branch becomes noisier.

Standard DDPM updates each branch from (x_t, z_u) with exact conditional means and fixed variances $\sigma^2_{x,t} = \beta_t$ and $\sigma^2_{z,t} = \beta_u$. This introduces two errors, both measured by the information terms of trajectory cost. First, a fixed variance ignores the information the other branch supplies, so it exceeds the conditional covariance by Eqs. (A.93)–(A.94). Second, the two branches are updated independently, which drops the dependence in Eq. (A.95). The two errors sit at opposite endpoints. At $s = 0$, the image always sees the clean representation, so its variance excess is largest, while no step updates both branches. At $s = T$, the image sees a noisy representation, so its excess is smallest, while every step drops dependence. Trajectory cost is minimized between these endpoints, and the final image KL below shows that the coupling it selects improves generation over both.

Final image KL. Under Assumption A.3, conditional means and probability-flow velocities are linear in the conditioning variables, so every sampler step is linear with additive Gaussian noise. We write $\hat{x}_t$ and $\hat{z}_u$ for generated states, to distinguish them from the forward variables, and $\hat{\Sigma}$ for the covariance of the current generated pair. Sampling starts from $(\hat{x}_T, \hat{z}_T) \sim \mathcal{N}(0, I_{d+k})$, so all generated states remain zero-mean Gaussian. A coupled step maps $(\hat{x}_t, \hat{z}_u)$ to $(\hat{x}_{t-1}, \hat{z}_v)$ with $u = \tau_s(t)$ and $v = \tau_s(t-1)$:

$$\begin{pmatrix} \hat{x}_{t-1} \\ \hat{z}_v \end{pmatrix} = A_t \begin{pmatrix} \hat{x}_t \\ \hat{z}_u \end{pmatrix} + \begin{pmatrix} \sigma_{x,t}\, \eta^x_t \\ \sigma_{z,t}\, \eta^z_t \end{pmatrix}, \qquad \hat{\Sigma} \leftarrow A_t \hat{\Sigma} A_t^\mathsf{T} + \mathrm{diag}(\sigma^2_{x,t} I_d, \sigma^2_{z,t} I_k), \qquad \text{(A.96)}$$

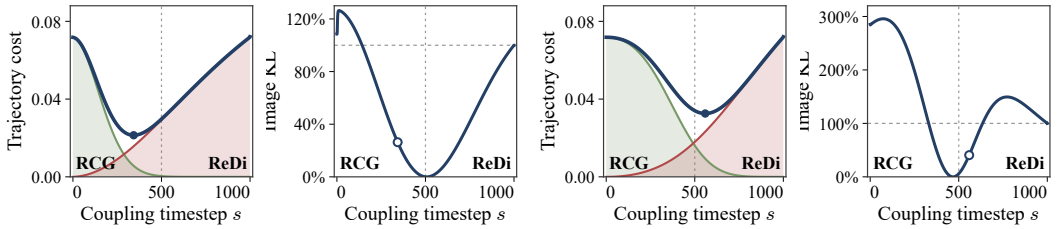
where η^x_t, η^z_t are independent standard Gaussians. For DDPM, the rows of A_t are the conditional-mean coefficients of x_{t-1} and z_v given (x_t, z_u); for Heun, A_t is one Heun step of the linear velocity field. Both A_t and the noise scales depend on s through u and v, and the representation row is the identity when $u = v$. Representation-only steps before coupling update $\hat{z}_u$ to $\hat{z}_{u-1}$ in the same form and hold $\hat{x}_T$. The generated image is $q_s(x_0) = \mathcal{N}(0, S_s)$, where S_s is the image block of the final $\hat{\Sigma}$. Since $p(x_0) = \mathcal{N}(0, I_d)$,

$$\mathrm{KL}\big(p(x_0) \,\|\, q_s(x_0)\big) = \tfrac{1}{2}\big[\mathrm{tr}(S_s^{-1}) - d + \log\det S_s\big]. \qquad \text{(A.97)}$$

DDPM sampling. DDPM uses $\sigma^2_{x,t} = \beta_t$ and $\sigma^2_{z,t} = \beta_u$, set to zero for a held representation or a clean output (Ho et al., 2020). Figure A.1(a) uses the Gaussian setting in Sec. 4.4 and the closed forms in Appendix A.7 ($\rho = 0.5$, $T = 1000$, linear $\beta_t \in [10^{-4}, 0.02]$). The interior coupling selected by trajectory cost reduces image KL by 73.7% relative to the better endpoint, and JERI's midpoint reduces it by 99.96%.

PF-ODE sampling. The linear path $x_t = (1 - t/T)x_0 + (t/T)\varepsilon^x$, and analogously z_t, has probability-flow velocity $\mathrm{d}x_t/\mathrm{d}t = T^{-1}\mathbb{E}[\varepsilon^x - x_0 \mid x_t, z_u]$ (Song et al., 2021; Lipman et al., 2023). Heun's method (Karras et al., 2022), including the final corrector, gives a linear A_t in Eq. (A.96) with $\sigma_{x,t} = \sigma_{z,t} = 0$. Trajectory cost uses the Markov forward process in Assumption A.1 with the same

normalized marginals, $\bar{\alpha}_t = (T-t)^2/[(T-t)^2+t^2]$, taking $\bar{\alpha}_T = 0$ by continuity. With the same ρ, T, and coupling schedule, Fig. A.1(b) shows that the interior coupling selected by trajectory cost reduces image KL by 59.4% relative to the better endpoint, and JERI's midpoint reduces it by 94.2%.



(a) DDPM (b) Heun

Figure A.1: Trajectory cost and image KL for (a) DDPM with $\sigma_t^2 = \beta_t$ and (b) Heun sampling of the linear-path probability-flow ODE. Blue denotes trajectory cost; green and red show its imageless and noisy-representation components. Circles mark the coupling selected by trajectory cost.

For both samplers, the final image KL is minimized in the interior, and the coupling selected by trajectory cost improves it over both endpoints. In this Gaussian setting, trajectory cost, the criterion we probe on ImageNet in Sec. 4.4, thus selects couplings that improve the final image distribution of standard samplers.

B Probing with Real Images on ImageNet

B.1 Probing with Forward-Process Information

Theorem 4.1 is stated in terms of mutual-information quantities of the forward process. We therefore test these quantities directly, without training a neural network. Since full high-dimensional mutual information cannot be reliably estimated, we reduce the image latent and representation to matched scalar coordinates and estimate mutual information on these one-dimensional variables.

On the image side, we use the VAE latent x_0; on the representation side, we use a frozen SSL representation z_0 from DINOv2 or MoCo v3. The two spaces share no coordinate system, so a direction in one has no counterpart in the other. We match them by the class index: we train two linear classifiers on ImageNet, one mapping x_0 to the 1000 classes and one mapping z_0 to the 1000 classes. Both see only clean states, share no parameters, and are frozen afterwards; the classification objective is used only to fix their weights. For class c, let $w_{x,c}$ and $w_{z,c}$ be its unit-norm weight vectors in the two classifiers, i.e. the evidence for class c in each space, learned separately. We define

$$X_t^c := w_{x,c}^\top x_t, \qquad Z_t^c := w_{z,c}^\top z_t, \tag{B.1}$$

the two logits of class c, up to a fixed scale and shift that leave mutual information unchanged.

Because $w_{x,c}$ and $w_{z,c}$ are unit-norm, $w^\top \epsilon$ is again a standard Gaussian, and linear projection commutes exactly with the forward process:

$$X_t^c = \sqrt{\bar{\alpha}_t}\, X_0^c + \sqrt{1-\bar{\alpha}_t}\, \epsilon_{t,c}^x, \qquad Z_t^c = \sqrt{\bar{\alpha}_t}\, Z_0^c + \sqrt{1-\bar{\alpha}_t}\, \epsilon_{t,c}^z, \tag{B.2}$$

where $\epsilon_{t,c}^x$ and $\epsilon_{t,c}^z$ are standard Gaussian noises. So (X_t^c) and (Z_t^c) are one-dimensional forward processes on the same schedule $\bar{\alpha}_t$, and Assumption A.1 holds for them verbatim.

B.2 Reconstructing the Trajectory Cost

Both costs are conditional mutual informations, which we do not estimate directly. Under Assumption A.1, they can be written as differences of pairwise mutual informations between the projected scalars:

$$I_t^z = I(X_t^c; Z_{t-1}^c) - I(X_t^c; Z_t^c), \tag{B.3}$$

$$I_{t,\tau}^x = \big[I(X_{t-1}^c; Z_0^c) - I(X_{t-1}^c; Z_\tau^c)\big] - \big[I(X_t^c; Z_0^c) - I(X_t^c; Z_\tau^c)\big]. \tag{B.4}$$

Table B.1: Measured sufficient conditions for Theorem 4.1. Each entry reports the probed timesteps where the corrected margin is positive.

Condition	Expression	DINOv2	MoCo v3
(C1), leaving RCG	$m_0(s) - \Delta(s) > 0$	$50-200$	$50-200$
(C2), leaving ReDi/REG	$m_T(s) - \Delta(s) > 0$	$200-950$	$200-950$

Every term is a mutual information between two scalars, estimated from ImageNet samples. We evaluate these on the timesteps of the diffusion process, pairing X_t^c with $Z^c_{\tau_s(t)}$, and substitute into the trajectory cost to obtain $\mathcal{C}^c_{\mathrm{JERI}}(s)$. At $s = 0$ the Noisy Representation Cost vanishes because $\tau_0(t) = 0$; at $s = T$ the Imageless Cost is empty. The reported cost is the average of $\mathcal{C}^c_{\mathrm{JERI}}(s)$ over the 1000 classes.

B.3 Measured Trajectory Cost and Sufficient Conditions

Testing the sufficient conditions. The two conditions in Theorem 4.1 need not hold at the same timestep. Condition (C1) checks whether the cost removed by leaving RCG still exceeds the newly introduced cost after the warped pairing correction. Condition (C2) gives the corresponding comparison when leaving ReDi/REG. Table B.1 reports the probed timesteps where the corrected margins are positive.

Thus the measured quantities verify both sufficient conditions for each representation space, with the actual two-time pairing used by $\mathcal{C}_{\mathrm{JERI}}$, and the two ranges are identical across the two representation spaces.

C Representation Alignment in Unconditional Generation

We conduct an additional experiment on unconditional generation with REPA(Yu et al., 2025) to reveal distinct properties between conditional and unconditional generation.

We select REPA because it explicitly aligns intermediate DiT features with DINOv2 representations, achieving significant improvements in both convergence speed and generation quality. Since these benefits are hypothesized to arise from feature alignment rather than class label information, REPA's effectiveness would be expected to extend to unconditional generation.

To validate this, we train DiT-B with and without REPA using our ImageNet subset setting (Sec. 6.1), following REPA's recommended hyperparameters to ensure fair comparison. Surprisingly, we observe notably different behavior compared to the reported class-conditional results, as shown in Tab. C.1: REPA in its straightforward format does not lead to improved unconditional generation. While a thorough analysis of REPA for unconditional generation requires further investigation, this observation suggests that conditional and unconditional generation exhibit distinct properties. Given that class labels are the primary difference between our setting and REPA's original experiments, we hypothesize that class labels significantly influence representation-image coherence, which motivates our choice of unconditional generation as a cleaner testbed.

Table C.1: On the ImageNet-subset unconditional setting, adding REPA to DiT-B reduces sample quality relative to the baseline. Contrary to REPA's effectiveness for conditional generation, REPA does not directly improve unconditional generation, suggesting that class labels influence representation-image coherence.

Method	FID ↓	IS ↑
DiT	75.23	25.01
DiT w/ REPA	80.10	21.11

D Implementation Details

In Sec. 6.1, we briefly discuss the implementation details of our framework. This section provides further details comprehensively.

D.1 Network Architecture

Our architecture follows RCG (Li et al., 2024a) but extends it with bi-directional conditioning so that image and representation states can co-evolve. Concretely, we train two branches jointly: (i) an *image generation branch* parameterized by a DiT backbone, and (ii) a *representation synthesis branch* parameterized by an MLP.

Image Generation. Unless stated otherwise, we use DiT-B and DiT-XL (Peebles & Xie, 2023) as our default image generators. DiT-B has 130M trainable parameters, with a hidden dimension of 768 and a conditioning embedding dimension of 256. Following RCG, we replace the original class label conditioning with representation. At diffusion step t, the image generation branch predicts the noise for x_t conditioned on the representation branch. During training, z_t is obtained by applying the DDPM (Ho et al., 2020) forward process to the clean representation z_0 from the image encoder. Compared to RCG, which conditions on the fixed z_0 for all timesteps, our image generation branch consumes the time-varying z_t as its conditioning signal.

Representation Generation. We use a 16-layer MLP with 1536 hidden units per layer as our representation diffusion model, sharing the same architecture as RCG. The representation target space is the MoCov3 ViT-B embedding space (Chen et al., 2021) without per-sample standardization. At diffusion step t, the representation synthesis branch receives the noisy state z_t together with the contexts from the image generation branch (projected intermediate DiT features, as explained in Sec. 5.2 and below) to predict z_{t-1}.

Conditioning from Images to Representations. Recent work (Wang et al., 2025b; Leng et al., 2025; Yu et al., 2025) suggests that DiT exhibits an implicit encoder–decoder structure, with earlier blocks behaving like an encoder and later blocks like a decoder. Guided by these observations, we follow the setup in (Wang et al., 2025b) and treat layers 8 and 22 in DiT-B and DiT-XL as encoder layers, respectively. As described in Sec. 5.2, we retrieve the intermediate features at the corresponding encoder layer and project them into the representation branch. We keep the original RDM design and map the image features into a 512-dimensional condition embedding.

Our ablations in Sec. 6.3 show that teacher forcing with stop-gradient (Sec. 5.3) stabilizes the co-training dynamics of the two branches. Concretely, the image features are detached before being projected and fed into the representation branch. Symmetrically, the representation condition used in the image branch is obtained by forward-noising the clean SSL representation z_0 at each timestep.

D.2 Utilizing DINOv2 Representations

Since DINOv2 has a much larger dimensionality of 768, we follow the practice of ReDi (Kouzelis et al., 2025) by applying principal component analysis (PCA) to DINOv2's representation and compressing the feature dimension to 256, identical to MoCov3 (Chen et al., 2021) and the representation diffusion model of our RCG baseline. As discussed in DLC (Lavoie et al., 2025), the feature space of DINOv2 is significantly more challenging to synthesize than MoCov3 (Chen et al., 2021), making RCG's marginalized representation generation incapable of fitting. So we adopt the per-sample standardization discussed in (Li et al., 2024a) by default to decrease the challenges of DINOv2 synthesis. All other training settings are explained below.

D.3 Training

Loss Function. Our loss function reflects the learning objective of jointly modeling the evolution of representations and images, as indicated by Sec. 4.3 and Sec. 5.2. It also follows our proposed partially joint modeling (Sec. 5.1, Fig. 5b) to maintain better coherence between representations and images.

Let $\epsilon_x^{\theta_x}$ and $\epsilon_z^{\theta_z}$ denote the image and representation diffusion models, respectively. As in Sec. 5.1, our partially joint modeling objective involves two stages: (1) before timestep $t = s$, the representation branch denoises independently; (2) afterwards, the representation and image branches utilize bi-directional conditioning (Sec. 5.2) to jointly evolve. Accordingly, our loss function contains three terms: (1) for $s \leq t \leq T$, a denoising term for the representation branch; (2) for $0 \leq t \leq s$, two denoising terms for representation and image generation, respectively. For partially joint modeling, the image at timestep t forms a pair with the representation at $[\frac{st}{T}]$, as explained in Eqn. 5. Formally,

the loss function is:

$$\begin{aligned}\mathcal{L}(\theta) = w_1\, \mathbb{E}_{s\le t'\le T} \left\|\epsilon'_z - \epsilon_z^{\theta_z}\left(z_{t'}, x_\phi, t'\right)\right\|_2^2 \\ + w_2\, \mathbb{E}_{0<t\le T} \left\|\epsilon_x - \epsilon_x^{\theta_x}\left(x_t, z_{[\frac{st}{T}]}, t\right)\right\|_2^2 \\ + w_3\, \mathbb{E}_{0<t\le T} \left\|\epsilon_z - \epsilon_z^{\theta_z}\left(z_{[\frac{st}{T}]}, x_t, [\frac{st}{T}]\right)\right\|_2^2\end{aligned} \tag{D.1}$$

In Eqn. D.1, the first term denotes the independent representation at the early steps, where x_ϕ is a learnable empty image condition. The last two terms correspond to the joint image and representation synthesis stage. We use $w_1 = w_3 = 0.5$ and $w_2 = 1$.

Training Settings. We follow DiT's setting (Peebles & Xie, 2023), which trains both representation and image generation models with a constant learning rate of 1e-4 without weight decay. This setting is also adopted in RCG (Li et al., 2024a) 's open-source code. For the training on full ImageNet, we follow the same settings of ReDi (Kouzelis et al., 2025) for a fair comparison, including the batch size, optimizer, and training step number. Specifically, we use AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2017) optimizer and a batch size of 256. To adopt classifier-free guidance (Ho & Salimans, 2022), we randomly drop the image and representation condition by 10% in the bi-directional conditioning framework, replaced with learnable null conditions. During training, we follow the common diffusion setup (Ho et al., 2020) by discretizing the timesteps into $T = 1000$. As for our partially joint modeling, we set $s = 500$ by default, and analyze different values in Sec. 6.4.

D.4 Inference

Sampling. As described in Sec. 5.1, we utilize the partially joint conditioning and use the following two stages of sampling. (1) In the first stage, the representation branch denoises independently without the image conditions. (2) In the second stage, the two branches of representation and images co-evolve under the bi-directional conditioning described in Sec. 5.2, where x_t and $z_{[st/T]}$ form a pair, so that both the image and representation branches achieve the clean state of $t = 0$ simultaneously. Concretely, we employ 200 steps for the sampling of z_0 and 100 steps for the sampling of x_0, where bi-directional conditioning becomes active after the representation is independently sampled for half of the steps, corresponding to our hyper-parameter of $s = 500$.

Classifier-free Guidance. Our bi-directional conditioning between the image and representation branches naturally supports classifier-free guidance (CFG) across the two generators. For the simplicity of notations, we denote $t' = [\frac{st}{T}]$, and derive the CFG for bi-directional conditioning below:

$$\begin{aligned}\tilde{\epsilon}_z^{\theta_z}(z_{t'}, x_t, t') = \epsilon_z^{\theta_z}(z_{t'}, x_\phi, t') \\ + w_z\Big(\epsilon_z^{\theta_z}(z_{t'}, x_t, t') - \epsilon_z^{\theta_z}(z_{t'}, x_\phi, t')\Big),\end{aligned} \tag{D.2}$$

$$\begin{aligned}\tilde{\epsilon}_x^{\theta_x}(x_t, z_{t'}, t) = \epsilon_x^{\theta_x}(x_t, z_\phi, t) \\ + w_x\Big(\epsilon_x^{\theta_x}(x_t, z_{t'}, t) - \epsilon_x^{\theta_x}(x_t, z_\phi, t)\Big),\end{aligned} \tag{D.3}$$

where $\tilde{\epsilon}_x^{\theta_x}(x_t, z_{t'}, t)$ and $\tilde{\epsilon}_z^{\theta_z}(z_{t'}, x_t, t')$ are the results after applying CFG, and z_ϕ and x_ϕ are learnable null conditions as explained in Sec. D.3. We also adopt the linear CFG formulation (increasing the CFG weights from 1 to maximum linearly during the diffusion steps) and search for the optimal guidance weight for both our JERI and RCG for a fair comparison. Unless otherwise specified, all reported quantitative metrics are computed using the corresponding best-performing CFG weights for optimal image FID.

D.5 Selection of Classes with Different Diversities

In Sec. 6.5.1, we experiment with different sets of class diversities from ImageNet (Deng et al., 2009) to analyze the behaviors of our JERI and RCG (Li et al., 2024a). The three sets follow the same training setup; only the data differs. This section explains how the classes are selected to follow the trend from least diverse to most diverse.

For the least diverse classes, we notice the proximity of ImageNet classes, *e.g.*, classes with similar semantic meanings are generally close in indexes. Therefore, we utilize the first 100 classes from

ImageNet in our Table 4. For example, this set of classes covers multiple types of fish and snakes, instead of following a uniform distribution in semantics.

For the random classes, we utilize the 100 classes selected from the subsets of timm's ImageNet (timm), following the same setup of our main comparison (Table 5). Therefore, the set of random classes mimics the distribution of regular ImageNet.

For the most diverse classes, we first calculate the mean MoCov3 (Chen et al., 2021) embeddings for the 1000 classes in ImageNet, then utilize the Gonzalez algorithm (Gonzalez, 1985) to find the 100 classes with the furthest distance. Intuitively, the Gonzalez algorithm adopts a greedy iterative approach that selects the point farthest from the current set of chosen centroids. At each step, the algorithm adds the class that maximizes the minimum distance to any already selected class. This strategy forces the selection to span the extremes of the embedding space and achieve maximal diversity.

D.6 Efficiency Analysis

Table D.1 shows that our model JERI adds marginal or no overhead compared with the baselines and has similar efficiency. These results suggest that our improvement comes from the coupling between representations and images.

Table D.1: Efficiency comparison of ReDi, RCG, and JERI.

Metric	ReDi	RCG	JERI
Image branch params	675.0M	674.3M	675.1M
Rep branch params	—	131.5M	131.5M
Total params	675.0M	805.8M	806.7M
Combined FLOPs	237.27G/forward	237.51G/forward	237.51G/forward
Training image forward/backward	1 joint	1 image	1 image
Training rep forward/backward	—	1	1
Image/joint sampling steps	100	100	100
Rep sampling steps	—	200	200
Training peak memory	41.6 GB	44.4 GB	44.4 GB
Inference peak memory	14.4 GB	12.2 GB	12.2 GB
Training latency	0.22 s/iter	0.25 s/iter	0.25 s/iter
Sampling latency	0.65 s/image	0.35 s/image	0.36 s/image

Parameter Count. ReDi jointly predicts image latents and representations within one DiT, using additional representation input embeddings and prediction heads but no separate representation denoiser. RCG and JERI instead use a separate 131.5M-parameter representation denoiser in addition to the DiT-XL/2 image backbone.

Although this gives JERI 19.5% more parameters than ReDi, the representation branch is an MLP and contributes only 0.24 GFLOPs, which is well below that of the image DiT branch. Relative to RCG, JERI additionally introduces lightweight projection heads for image-to-representation conditioning, with only 0.86M parameters and 0.001 GFLOP.

Training Efficiency. ReDi performs one joint DiT forward/backward pass per iteration. RCG and JERI perform one image-DiT and one lightweight representation-denoiser forward/backward pass. Although Eq. (D.1) contains two representation losses, they are computed from the disjoint pre-coupling and coupled subsets within the same batched representation pass.

We measure all methods on one H100 GPU with batch size 64. JERI has the same measured cost as RCG, showing that its image-to-representation conditioning introduces no measurable overhead at the reported precision. ReDi is approximately 12% faster per iteration and uses approximately 6% less peak memory because it does not use a separate representation denoiser.

Inference Efficiency. We measure all methods on one H100 GPU under the same sampling setup. ReDi jointly predicts images and representations within each DiT step, whereas RCG and JERI use a lightweight separate representation denoiser.



Figure D.1: Additional visualizations of JERI.

E Qualitative Results

We provide more qualitative visualization, in addition to the results in Sec. 6.6. As shown in Fig. D.1, our model can generate reasonable images. However, we also observe similar issues of generation at the scale of models smaller than DiT-XL (Peebles & Xie, 2023), with limited parameters, that exhibit artifacts or unrealistic appearances.

F Limitations, Future Works, and Societal Impacts

Limitations and Future Work. Our JERI is a general framework that models representations and images jointly. We are constrained by computational resources, and potential future work includes verifying our proposed solution at larger scales and in multi-modal scenarios, such as editing domains. Therefore, we expect to extend our theoretical intuitions to effectively enhance more models and to the domains of text-to-image generation.

Impact Statement

As a general methodology investigating the synergy between images and representations, our JERI does not specifically introduce negative societal impacts. However, our JERI unavoidably reflects the bias of the internet data, similar to other generative models. From this perspective, our major focus is more towards the understanding of representation-image coupling, which may alleviate this issue.