REJECT TO ALIGN: SELECTIVE SAMPLE REJECTION FOR REPA

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation alignment (REPA) has become a standard technique for accelerating Diffusion Transformer (DiT) training by regularising hidden states toward features from a pretrained encoder. However, REPA computes its alignment gradient as a plain arithmetic mean over the batch, assigning equal weight to every sample regardless of alignment quality. We use a variational analysis to relate the population REPA objective to a Barber–Agakov lower bound on the conditional mutual information between DiT and encoder features. This analysis interprets cosine similarity as a positive-pair compatibility score and motivates examining per-sample alignment gradients. The gradient analysis characterises how alignment updates vary with cosine similarity. Complementary empirical measurements link rejected samples to weaker alignment and a higher alignment-to-denoising gradient-norm ratio. These observations motivate using large per-sample loss deviations as a practical indicator of atypical alignment updates. Based on this analysis, we propose Reject-REPA (R^2EPA), a selective sample rejection mechanism that filters alignment-loss outliers via a σ-based threshold grounded in the Vysochanskij–Petunin inequality, a dual detector combining Z-score + MAD for outlier identification, and cross-batch EMA statistics for adaptive thresholding. R^2EPA introduces no additional model parameters and serves as a drop-in replacement for standard REPA training. Experiments on different generation tasks suggest that selective alignment can improve generation quality in the evaluated settings.

1 INTRODUCTION

Diffusion Transformers (DiTs) (Peebles & Xie, 2022; Ma et al., 2024) have replaced the previous U-Net backbone (Karras et al., 2022; Dhariwal & Nichol, 2021; Rombach et al., 2022) as the dominant architecture for high-fidelity image generation (Park et al., 2026; Lee et al., 2026), owing to their scalability and capacity to model complex data distributions. Many recent works (Yu et al., 2025; Singh et al., 2025; Wang et al., 2025b; Wu et al., 2025) have demonstrated that the intermediate representations play an important role in both training efficiency and generation quality, establishing *representation alignment (REPA)*, the practice of regularising DiT hidden states toward features from a powerful pretrained encoder (e.g., DINO (Oquab et al., 2023), MAE (He et al., 2021)), as a broadly effective training accelerator for modern DiT pipelines.

REPA (Yu et al., 2025) formalises this intuition by adding a patch-level cosine similarity alignment term to the denoising objective, improving training efficiency and generation quality. The simplicity and effectiveness of REPA have inspired a series of subsequent improvement efforts: iREPA (Singh et al., 2025) replaces the Multilayer Perceptron (MLP) projection head with Convolutional Neural Networks (CNN) to provide better spatial structure, REPA-E (Leng et al., 2025) incorporates encoder fine-tuning to close the domain gap, and REG (Wu et al., 2025) introduces an additional cls-token for richer alignment. In standard REPA, the alignment gradient is computed as a plain arithmetic mean over the batch, assigning equal weight to each sample irrespective of its alignment quality.

Recently, HASTE (Wang et al., 2025b) provides evidence that this uniform treatment is problematic, as it observes that in later training iterations, the REPA gradient and the denoising gradient can conflict, causing alignment to actively degrade performance. HASTE addresses this by early-stopping the alignment objective. While HASTE identifies a temporal conflict between REPA and denoising gradients, it overlooks a finer-grained, per-sample perspective: not all samples within a batch

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.

contribute equally at any given step. This leaves open a natural question: *can we improve alignment training by selectively adjusting individual samples' contributions?*

To answer this question, we first ask a more fundamental one: when is a per-sample alignment gradient informative? Section 3.2 connects the population REPA objective to a Barber–Agakov lower bound on the conditional mutual information (Cover, 1999; Barber & Agakov, 2003) $I(Z_1; Z_2 \mid T)$ between DiT and encoder features (Proposition 1). Under this view, $-\ell_i$ serves as a sample-wise positive-pair compatibility score in the empirical variational objective, making score heterogeneity a natural lens for analysing per-sample gradient geometry.

The gradient analysis shows that weak or orthogonal alignment ($\mathrm{sim}_i \approx 0$) produces the largest feature-space residual; after propagation through the sample-dependent network Jacobian, the resulting parameter gradient may conflict with the aggregate alignment direction (Case A, Proposition 2). At the endpoints $\mathrm{sim}_i \approx \pm 1$, the cosine gradient instead saturates: positive alignment corresponds to a low-loss, low-utility update (Case B, Proposition 3), while antipodal alignment yields high loss with a similarly vanishing residual. These regimes motivate treating a large deviation $|\ell_i - \bar{\mu}|$ as a practical indicator of atypical alignment updates, whose usefulness we examine empirically.

Our empirical analysis complements this geometric view: at a standard REPA model, samples selected for rejection have lower alignment similarity and a larger alignment-to-denoising gradient-norm ratio within the high-noise stratum (Section 5.3). These observations motivate selectively moderating atypical alignment contributions. Accordingly, we propose Reject-REPA (R^2EPA), a selective sample rejection mechanism for REPA training. Leveraging the empirical observation (Figure 2) that the per-sample REPA loss is approximately unimodal after warm-up, we use the Vysochanskij–Petunin inequality (Pukelsheim, 1994) to calibrate a σ-based rejection threshold. We design a dual-detector mechanism (Z-score + MAD with OR fusion) to identify outliers when extreme losses may contaminate batch statistics. Boundary tracking in Figure 3 illustrates how the two branches contribute to the effective rejection thresholds at different training steps. To balance historical trends with current-batch information, we maintain cross-batch exponential moving average (EMA) statistics for adaptive thresholding. R^2EPA can be integrated into REPA without additional model parameters. Unlike HASTE, which stops alignment globally, R^2EPA preserves the objective while filtering only statistically atypical per-sample terms. Our experiments on ImageNet and COCO suggest that selective alignment can benefit generation quality. Our main contributions are as follows.

- **Analysis.** We use a variational analysis to relate REPA to conditional information maximisation (Proposition 1) and examine how its alignment gradient varies across cosine-similarity regimes (Propositions 2 and 3). This analysis motivates per-sample loss deviation as a practical indicator for identifying atypical alignment updates.
- **Method Design.** We propose R^2EPA, a simple drop-in replacement for REPA that selectively rejects outlier samples via a statistically grounded dual-detector mechanism, requiring no architectural modification.
- **Empirical Evaluation.** We study R^2EPA on ImageNet and COCO, reporting generation metrics and training trajectories. Offline gradient measurements provide additional context for the samples selected by the detector. We also examine detector behaviour and compare the Case A and Case B rejection branches separately and together.

2 Related Work

Improved Diffusion Model Training. Denoising diffusion probabilistic models (Ho et al., 2020) optimize a denoising score-matching objective (Song et al., 2021) across all noise levels. Subsequent work has refined this training target along several dimensions. A prominent line of research addresses the non-uniformity of the diffusion process, wherein different noise levels contribute unequally to the overall training objective, causing naive uniform sampling to result in suboptimal computational efficiency. To mitigate this, a variety of methods have been proposed that adopt timestep re-sampling (Zheng et al., 2024; Kim et al., 2025a;b), loss re-weighting strategies (Choi et al., 2022; Go et al., 2024; Hang et al., 2023; Yu et al., 2024), and both (Zheng et al., 2025; Wang et al., 2025a), yielding substantial improvements in the training efficiency of diffusion models. On the architecture side, recent works such as DiT (Peebles & Xie, 2022), SiT (Ma et al., 2024) and JiT (Li & He, 2025) replace the prevalent U-Net backbone (Karras et al., 2022; Rombach et al., 2022) with

Transformer-based architectures, establishing a new paradigm in which the quality of intermediate representations plays a decisive role in both training efficiency and generation quality.

Representation Alignment for Diffusion Models. To obtain a better feature representation, REPA (Yu et al., 2025) introduces patch-level cosine alignment between DiT hidden states and a frozen vision encoder (e.g., DINOv2 (Oquab et al., 2023)), substantially accelerating convergence without modifying the sampling procedure. Subsequent work has pushed this paradigm further: iREPA (Singh et al., 2025) extends the projection head from MLP to CNN. REPA-E (Leng et al., 2025) incorporates encoder fine-tuning to close the domain gap between generation and reconstruction features. REG (Wu et al., 2025) designs an additional cls-token for richer alignment. Together, these methods demonstrate that representation alignment is a broadly effective training accelerator and have established it as a standard component of modern DiT pipelines. A property shared by all of the above methods is that the alignment gradient is computed as a plain arithmetic mean over all batch samples, assigning equal weight irrespective of alignment quality. Our analysis and experiments suggest that this uniform treatment can be suboptimal when a batch contains atypical alignment updates. HASTE (Wang et al., 2025b) addresses a related temporal issue by stopping alignment globally, but does not identify alignment-loss outliers within a batch. We instead investigate a cross-batch statistical strategy for selectively filtering such outliers.

3 Theoretical Analysis of Representation Alignment

3.1 Preliminary

Flow Matching. Let $x_0 \in p_{\text{data}}$ be a training sample and $t \sim \mathcal{U}[0,1]$ a diffusion timestep. The flow-matching forward process is defined as $x_t = (1-t)x_0 + t\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$. In the training stage, we train a DiT model with parameters θ based on the following loss

$$\mathcal{L}_{velocity}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, x_0\sim p_{\text{data}},\, \varepsilon\sim\mathcal{N}(0,I)} \left[\|v_\theta(x_t, t) - (\varepsilon - x_0)\|^2 \right], \tag{1}$$

where v_θ is the velocity field parameterized by θ, and the target vector field $u_t(x_t \mid x_0) = \varepsilon - x_0$ is the conditional flow that transports data samples x_0 toward Gaussian noise ε along a straight-line path. At the inference stage, generation proceeds by integrating the learned velocity field v_θ via the reverse SDE or ODE, *i.e.*, from $t = 1$ to $t = 0$. Starting from a pure noise sample $x_1 \sim \mathcal{N}(0, I)$, the trajectory is gradually denoised toward the data distribution p_{data}, yielding the generated sample $\hat{x}_0$.

Representation Alignment. In the training stage, the DiT with parameters θ produces hidden states $H_\theta^{(l)}(x_t, t) \in \mathbb{R}^{N\times d}$, and an MLP head $g_\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ maps them into the encoder feature space. Let $z_{1,i}^{[n]}$ and $z_{2,i}^{[n]}$ denote the ℓ_2-normalized patch-n features of the projected DiT hidden state and the frozen encoder output for sample i. REPA (Yu et al., 2025) augments the denoising objective with an additional patch-level cosine alignment regular term:

$$\mathcal{L} = \mathcal{L}_{\text{velocity}} + \lambda\, \mathcal{L}_{\text{REPA}}, \qquad \mathcal{L}_{\text{REPA}} = \frac{1}{B} \sum_{i=1}^{B} \ell_i, \tag{2}$$

where the *per-sample* alignment loss is defined as $\ell_i \triangleq -\frac{1}{N} \sum_{n=1}^{N} z_{1,i}^{[n]\top} z_{2,i}^{[n]} \in [-1, 1]$, and we denote the batch mean as $\bar{\mu} \triangleq \frac{1}{B} \sum_{i=1}^{B} \ell_i = \mathcal{L}_{\text{REPA}}$. Since $\mathcal{L}_{\text{REPA}}$ takes the form of an arithmetic mean over the batch, its gradient reduces to a simple average

$$\nabla_\theta \mathcal{L}_{\text{REPA}} = \frac{1}{B} \sum_i g_i, \tag{3}$$

where $g_i \triangleq \nabla_\theta \ell_i$. Consequently, every sample receives equal weight $1/B$ regardless of its alignment quality in standard REPA training.

3.2 A Variational Analysis of REPA

We use a variational analysis of cosine-based representation alignment to relate REPA to conditional information maximisation. Let $T \sim \mathcal{U}[0,1]$ denote the diffusion timestep random variable, with

t denoting its realisation. Let J be a patch index sampled uniformly from $\{1, \ldots, N\}$, and define $Z_1 = Z_1^{[J]}$ and $Z_2 = Z_2^{[J]}$ as the corresponding normalised DiT and frozen-encoder feature random variables. Consider the von Mises–Fisher (vMF) variational conditional density (Fisher, 1953; Govindarajan et al., 2023) on the unit hypersphere,

$$q_\kappa(z_2 \mid z_1) = C_{d'}(\kappa) \exp\big(\kappa z_1^\top z_2\big), \qquad \kappa > 0. \tag{4}$$

The vMF family provides a tractable variational model on the unit sphere. Its log density is affine in cosine similarity and therefore recovers the score used by REPA for normalised features.

Proposition 1 (A variational view of REPA)**.** *For the variational conditional density in equation 4,*

$$I(Z_1; Z_2 \mid T) \;\geq\; C_\kappa - \kappa\, \mathbb{E}[\mathcal{L}_{\mathrm{REPA}}], \qquad C_\kappa \triangleq H(Z_2 \mid T) + \log C_{d'}(\kappa). \tag{5}$$

When the data distribution and frozen encoder are fixed and κ is held constant, C_κ is independent of the trainable parameters. Consequently, minimising the population REPA objective can be viewed as maximising this particular Barber–Agakov variational lower bound.

The proof in Appendix A.1 follows from the standard variational conditional entropy argument (Barber & Agakov, 2003). The result interprets $-\ell_i$ as a Monte Carlo term in the empirical variational log-likelihood, or equivalently as a sample-wise positive-pair compatibility score. This view highlights heterogeneity across alignment samples and motivates examining how unusually high or low compatibility scores affect the geometry of the alignment gradient.

3.3 Gradient Analysis

Let $\mathrm{sim}_i^{[n]} \triangleq z_{1,i}^{[n]\top} z_{2,i}^{[n]}$ denote the cosine similarity between the normalised features of patch n in sample i. The gradient g_i is driven by the patch-level residual $\delta_i^{[n]} \triangleq z_{2,i}^{[n]} - \mathrm{sim}_i^{[n]} \cdot z_{1,i}^{[n]}$, which is the component of $z_{2,i}^{[n]}$ orthogonal to $z_{1,i}^{[n]}$. This feature-space geometry is architecture independent, whereas the magnitude and direction of the resulting parameter gradient additionally depend on the projection and network Jacobians.

Proposition 2 (Potential Directional Conflict at Weak Alignment, Case A)**.** *When* $\mathrm{sim}_i^{[n]} \to 0$*,* $\delta_i^{[n]} \to z_{2,i}^{[n]}$ *and* $\left\|\delta_i^{[n]}\right\| \to 1$*. Thus weak or orthogonal alignment produces a full-scale feature space residual. After propagation through the sample-dependent Jacobian, the resulting parameter gradient may deviate from, or even conflict with, the aggregate alignment direction.*

Proposition 3 (Vanishing Alignment Residual, Case B)**.** *When* $\mathrm{sim}_i^{[n]} \to 1$ *for all aligned patches* n*,*

$$\left\|\delta_i^{[n]}\right\|^2 = 1 - \big(\mathrm{sim}_i^{[n]}\big)^2 = (1 - \mathrm{sim}_i^{[n]})(1 + \mathrm{sim}_i^{[n]}) \to 0, \tag{6}$$

so the alignment signal vanishes. If the relevant network Jacobians remain bounded and the feature norms before normalisation are bounded away from zero, the corresponding parameter gradient norm also approaches zero, indicating a diminishing contribution from the alignment update.

Equation 6 also reveals another upper-loss extreme: when $\mathrm{sim}_i^{[n]} \to -1$, the features are negatively aligned, and the residual again vanishes. Hence, alignment loss and gradient magnitude capture complementary properties. The residual is largest around zero similarity, while both endpoints of the cosine range are first-order saturated. These regimes motivate using a large $|\ell_i - \bar{\mu}|$ as a simple indicator of atypical alignment updates: the upper-loss tail covers weak-to-negative compatibility, whereas the lower-loss tail captures near-complete positive alignment. We evaluate loss deviation empirically as a practical proxy for gradient utility.

4 Methods

4.1 Methodology Design

The Truncated Estimator. The limiting cases in Section 3 suggest that per-sample loss deviation $|\ell_i - \tilde{\mu}|$ can serve as a practical indicator of atypical alignment updates. In the upper tail, weak or

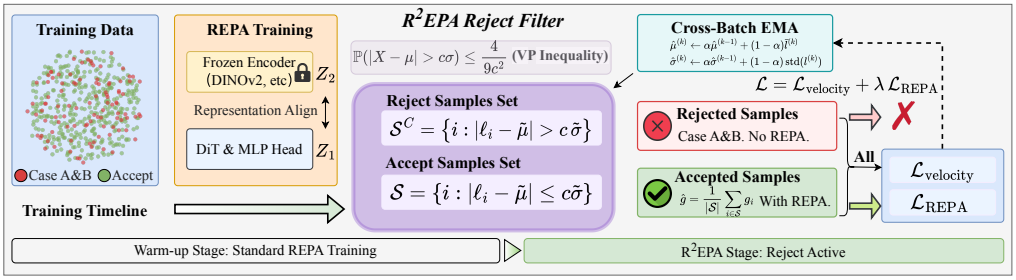

Figure 1: **The pipeline of R^{2}EPA.** The two-sided filter combines Z-score and MAD thresholds to identify candidate alignment-loss outliers, using cross-batch EMA statistics for the Z-score branch. Only accepted samples contribute to $\mathcal{L}_{\text{REPA}}$. Rejection is activated after a warm-up stage, once the loss distribution becomes approximately unimodal.

orthogonal compatibility can produce a large residual and a potentially conflicting update (Case A), while the antipodal limit combines maximal loss with a saturated residual. The lower tail corresponds to near-complete positive alignment and a vanishing alignment signal (Case B). This motivates evaluating a truncated REPA estimator that retains the *accepted set*

$$\mathcal{S}_{\text{valid}} = \{i : |\ell_i - \tilde{\mu}| \le c\,\tilde{\sigma}\}, \tag{7}$$

where $\tilde{\mu}$ and $\tilde{\sigma}$ are estimates of the mean and standard deviation of the per-sample losses, and $c > 0$ is a threshold parameter. The REPA gradient is then computed over this subset:

$$\hat{g} = \frac{1}{|\mathcal{S}_{\text{valid}}|} \sum_{i \in \mathcal{S}_{\text{valid}}} g_i. \tag{8}$$

The rule excludes samples whose alignment losses are unusually high or unusually low from the REPA update, while retaining all samples for the denoising objective. Table 4 evaluates the two rejection branches separately: Case A only rejects high-loss samples, whereas Case B only rejects low-loss samples. The offline measurements on a REPA model in Section 5.3 show that samples selected for rejection have lower average cosine similarity and a larger average ratio of the alignment gradient norm to the denoising gradient norm than retained samples.

Cross-batch Adaptive Statistics. To track the loss scale across batches, we interpolate the current-batch statistics with their exponential moving averages (EMA) at step k:

$$\tilde{\mu}^{(k)} = \beta\bar{\ell}^{(k)} + (1-\beta)\hat{\mu}^{(k-1)}, \qquad \tilde{\sigma}^{(k)} = \beta\,\text{std}(\ell^{(k)}) + (1-\beta)\hat{\sigma}^{(k-1)}. \tag{9}$$

We set $\beta = 0.5$. After detection, the EMA state is updated for the next step:

$$\hat{\mu}^{(k)} = \alpha\hat{\mu}^{(k-1)} + (1-\alpha)\bar{\ell}^{(k)}, \qquad \hat{\sigma}^{(k)} = \alpha\hat{\sigma}^{(k-1)} + (1-\alpha)\,\text{std}(\ell^{(k)}). \tag{10}$$

Threshold and Dual Detectors. The approximately unimodal loss distributions in Figure 2 motivate using the Vysochanskij–Petunin inequality (Pukelsheim, 1994) to calibrate the threshold scale c; Appendix A.2 gives the bound and discussion. We augment the Z-score rule with a median-based detector (Hampel, 1974; Leys et al., 2013). Let $\tilde{\ell} = \text{median}(\{\ell_i\})$ and $\text{MAD} = \text{median}(|\ell_i - \tilde{\ell}|)$. The MAD rule flags sample i when

$$\frac{0.6745\,|\ell_i - \tilde{\ell}|}{\text{MAD}} > c. \tag{11}$$

The consistency factor 0.6745 rescales the MAD to match the standard deviation under Gaussianity, so both detectors share the same threshold c. A sample is rejected if it is flagged by *either* detector (OR fusion), and $\mathcal{S}_{\text{valid}}$ retains the remainder. The standard-deviation and median-based branches respond differently to the loss distribution. Their OR fusion gives the effective boundaries at step k:

$$U^{(k)} = \min\left\{\tilde{\mu}^{(k)} + c\tilde{\sigma}^{(k)}, \tilde{\ell}^{(k)} + \frac{c\,\text{MAD}^{(k)}}{0.6745}\right\}, L^{(k)} = \max\left\{\tilde{\mu}^{(k)} - c\tilde{\sigma}^{(k)}, \tilde{\ell}^{(k)} - \frac{c\,\text{MAD}^{(k)}}{0.6745}\right\}.$$

A sample is retained when $L^{(k)} \le \ell_i \le U^{(k)}$. The bottom panel of Figure 3 summarises how often each branch determines these boundaries during the active phase. The branches complement each other in boundary selection: on each side, either MAD or Z-score can supply the tighter threshold as batch statistics change.

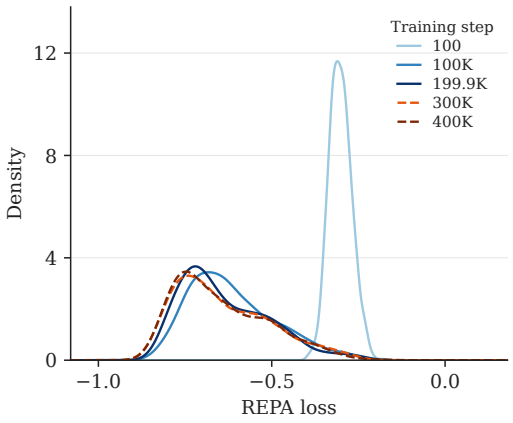

Figure 2: REPA loss distributions before (solid) and after (dashed) the 200K warm-up.

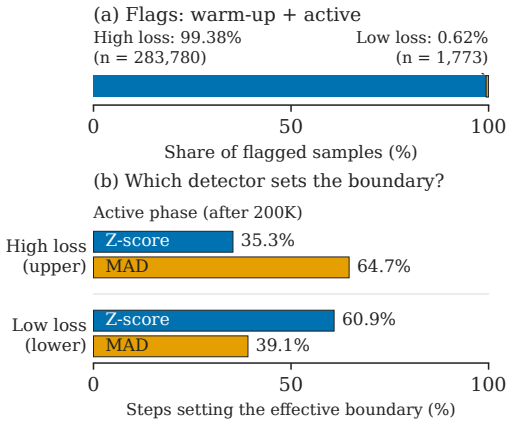

Figure 3: Loss-tail flags and detector contributions (Appendix B).

Algorithm 1 The Training Algorithm of R^{2}EPA.

Require: Batch $\{(x_0^{(i)}, t^{(i)})\}_{i=1}^{B}$; EMA statistics $\hat{\mu}, \hat{\sigma}$; threshold c; momentum α; interpolation weight $\beta = 0.5$; REPA weight λ; warmup steps k_0

Sample noise $\varepsilon^{(i)} \sim \mathcal{N}(0, I)$; compute noisy input $x_t^{(i)} = (1 - t^{(i)})x_0^{(i)} + t^{(i)}\varepsilon^{(i)}$
Compute velocity loss: $\mathcal{L}_{\text{velocity}}$ and per-sample REPA loss: $\ell_i = -\frac{1}{N}\sum_n z_{1,i}^{[n]\top} z_{2,i}^{[n]}$ for all i
if $k \le k_0$ **then** ▷ Warmup: No Rejection
    $\mathcal{S}_{\text{valid}} \leftarrow \{1, \ldots, B\}$
else ▷ R^{2}EPA: Rejecting Outlier Samples
    Form detection statistics: $\tilde{\mu} \leftarrow \beta\bar{\ell} + (1-\beta)\hat{\mu}, \quad \tilde{\sigma} \leftarrow \beta\,\mathrm{std}(\ell) + (1-\beta)\hat{\sigma}$
    Compute MAD statistics: $\tilde{\ell} \leftarrow \mathrm{median}(\ell), \quad \mathrm{MAD} \leftarrow \mathrm{median}(|\ell - \tilde{\ell}|)$
    Z-score detector: flag i if $|\ell_i - \tilde{\mu}|/\tilde{\sigma} > c$, MAD detector: flag i if $0.6745\,|\ell_i - \tilde{\ell}|/\mathrm{MAD} > c$
    $\mathcal{S}_{\text{valid}} \leftarrow \{i : i$ flagged by *neither* detector$\}$
end if
Update EMA: $\hat{\mu} \leftarrow \alpha\hat{\mu} + (1-\alpha)\bar{\ell}; \quad \hat{\sigma} \leftarrow \alpha\hat{\sigma} + (1-\alpha)\,\mathrm{std}(\ell)$
$\mathcal{L} \leftarrow \mathcal{L}_{\text{velocity}} + \lambda \cdot \frac{1}{|\mathcal{S}_{\text{valid}}|}\sum_{i \in \mathcal{S}_{\text{valid}}} \ell_i$ ▷ $\mathcal{L}_{\text{velocity}}$: All samples, No Rejection
Backpropagate $\mathcal{L}$; update θ, ϕ

Training Objective. All samples contribute during the first k_0 warm-up steps. Filtering then uses $\mathcal{S}_{\text{valid}} = \{i : L^{(k)} \le \ell_i \le U^{(k)}\}$ for the alignment, while the denoising term still uses the full batch:

$$\mathcal{L} = \mathcal{L}_{\text{velocity}} + \lambda \cdot \frac{1}{|\mathcal{S}_{\text{valid}}|} \sum_{i \in \mathcal{S}_{\text{valid}}} \ell_i. \tag{12}$$

The method requires no architectural changes. Algorithm 1 provides the complete procedure; Section 5.3 examines k_0 and c.

5 Experiments

5.1 Experimental Setup

Experimental Details. We follow the setting of REPA (Yu et al., 2025) as our baselines. Our main experiments use ImageNet (Deng et al., 2009), with an additional text-to-image evaluation on COCO (Lin et al., 2014). We adopt FID (Heusel et al., 2017), sFID (Nash et al., 2021), IS, Precision, and Recall (Kynkäänniemi et al., 2019) as our evaluation metrics. For the main ImageNet comparison, we use the SDE-250 sampler; alternative samplers are evaluated separately. More implementation details and settings are in Appendix C.

Table 1: Quantitative Comparison on ImageNet 256 × 256 at 400K Iterations w.r.t. Different Architectures. SiT-XL adopts interval CFG guidance (Kynkäänniemi et al., 2024) for both methods in the w/ CFG setting.

		w/o CFG					w/ CFG				
Model	**Method**	FID↓	sFID↓	IS↑	Prec↑	Rec↑	FID↓	sFID↓	IS↑	Prec↑	Rec↑
SiT-B	REPA	24.40	**6.4**	59.90	0.59	0.65	5.59	5.2	173.61	0.78	**0.54**
	R^{2}EPA	**22.12**	6.5	**66.99**	**0.60**	**0.65**	**5.39**	**5.2**	**177.15**	**0.79**	0.53
SiT-L	REPA	10.00	5.2	109.20	0.69	0.65	2.81	4.9	264.34	0.85	**0.55**
	R^{2}EPA	**9.50**	**5.2**	**114.65**	**0.69**	**0.66**	**2.74**	**4.8**	**266.63**	**0.85**	0.53
SiT-XL	REPA	7.90	5.1	122.60	0.70	0.65	1.93	4.6	281.89	0.82	0.59
	R^{2}EPA	**7.57**	**5.0**	**128.88**	**0.71**	**0.65**	**1.89**	**4.6**	**282.02**	**0.82**	**0.60**

5.2 Quantitative Comparison

Benchmark Analysis. Table 1 presents a quantitative comparison between R^{2}EPA and the REPA baseline across different model scales. For example, without CFG, FID changes from 24.40 to 22.12 on SiT-B and from 7.90 to 7.57 on SiT-XL. With standard CFG, SiT-L changes from 2.81 to 2.74; with interval CFG (Kynkäänniemi et al., 2024), SiT-XL changes from 1.93 to 1.89. The accompanying IS, precision, and recall results provide further context for these FID differences.

Table 2: R^{2}EPA on SiT-B (ImageNet 256 × 256, w/o CFG). (a) Training strategies and the iREPA setting. (b) Samplers and inference steps.

Method	FID↓	sFID↓	IS↑	Prec↑	Rec↑
REPA	24.40	6.4	59.90	0.59	0.65
Adaptive Non-Uni.	22.74	6.5	64.34	0.60	0.65
Min-SNR	22.60	**6.3**	66.84	0.59	0.65
HASTE	22.55	6.5	64.98	0.60	0.65
R^{2}EPA	**22.12**	6.5	**66.99**	**0.60**	**0.65**
iREPA	21.40	6.8	70.88	0.60	0.65
iREPA + R^{2}EPA	**20.50**	**6.7**	**72.34**	**0.60**	**0.66**

Sampler	FID↓	sFID↓	IS↑	Prec↑	Rec↑
REPA (ODE-100)	25.11	6.9	61.77	0.58	0.65
R^{2}EPA (ODE-100)	**24.47**	**6.8**	**63.13**	**0.58**	**0.65**
REPA (SDE-100)	24.29	8.0	63.98	0.59	0.65
R^{2}EPA (SDE-100)	**23.51**	**7.8**	**66.01**	**0.59**	**0.65**
REPA (SDE-250)	24.40	**6.4**	59.90	0.59	0.65
R^{2}EPA (SDE-250)	**22.12**	6.5	**66.99**	**0.60**	**0.65**

Further Comparisons and Analysis. In Table 2 (a), we compare R^{2}EPA with adaptive non-uniform weighting (Kim et al., 2025b;c), the early-stop strategy of HASTE (Wang et al., 2025b), and the static reweighting strategy of Min-SNR (Hang et al., 2023). On SiT-B, R^{2}EPA obtains FID 22.12, compared with 22.74 for adaptive non-uniform weighting, 22.55 for HASTE, and 22.60 for Min-SNR. Adding R^{2}EPA to the evaluated iREPA configuration (Singh et al., 2025) changes FID from 21.40 to 20.50. The method introduces no additional model parameters. Table 2 (b) reports the corresponding comparisons with ODE and SDE samplers at different inference steps.

More Comparison. Table 3 compares R^{2}EPA against representative diffusion models operating in pixel and latent spaces, including U-Net, hybrid, and Transformer architectures, on ImageNet 256 × 256. With interval CFG (Kynkäänniemi et al., 2024), R^{2}EPA obtains FID 1.62 and IS 306.2 at 200 training epochs, achieving the highest IS. On the text-to-image generation task on COCO, R^{2}EPA with SiT-B improves FID from 5.86 to 5.70 compared with the REPA baseline using the same architecture. This result suggests that the method also applies beyond class conditional generation. Appendix B further reports FDr^6 (Yang et al., 2026) as auxiliary evidence.

5.3 Analysis and Ablation

Rejection Behaviour. In Figure 3, high-loss observations account for 99.38% of pooled flags and low-loss observations for 0.62%. The active-phase boundary shares illustrate complementary detector roles: MAD sets the upper boundary at 64.7% of logged steps and Z-score at 35.3%; the lower-boundary shares are 39.1% and 60.9%, respectively. Thus, both branches contribute on each side at different steps. Appendix B gives the details.

Table 3: Comparison of different models on ImageNet 256×256 with classifier-free guidance. "$*$" denotes the use of interval CFG guidance (Kynkäänniemi et al., 2024). R^{2}EPA is evaluated at 200 training epochs. Best results across the entire table in **blue bold**.

Model	Epochs	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
Pixel diffusion						
ADM-U	400	3.94	6.1	186.7	0.82	0.52
VDM++	560	2.40	–	225.3	–	–
Simple diffusion	800	2.77	–	211.8	–	–
CDM	2160	4.88	–	158.7	–	–
Latent diffusion, U-Net						
LDM-4	200	3.60	–	247.7	**0.87**	0.48
Latent diffusion, Transformer + U-Net hybrid						
U-ViT-H/2	240	2.29	5.7	263.9	0.82	0.57
Latent diffusion, Transformer						
MaskDiT	1600	2.28	5.7	276.6	0.80	0.61
SD-DiT	480	3.23	–	–	–	–
DiT-XL/2	1400	2.27	4.6	278.2	0.83	0.57
SiT-XL/2	1400	2.06	4.5	270.3	0.82	0.59
+ REPA	200	1.96	4.5	264.0	0.82	0.60
+ R^{2}EPA (ours)	200	1.94	4.5	275.6	0.82	0.60
+ REPA*	800	**1.42**	4.7	305.7	0.80	**0.65**
+ R^{2}EPA* (ours)	**200**	1.62	**4.5**	**306.2**	0.81	0.62

Empirical Gradient Measurements. At a frozen REPA model, detector-selected samples in the high-noise stratum have lower mean cosine similarity (0.423 vs. 0.589) and a larger alignment-to-denoising gradient-norm ratio (3.82 vs. 2.80). More details are shown in Appendix B.

Table 4: Case A/B rejection on SiT-XL at 400K steps without CFG. Case A rejects higher-loss samples; Case B rejects lower-loss samples.

Method	FID↓	sFID↓	IS↑	Precision↑	Recall↑
REPA	7.90	5.1	122.60	0.70	0.65
R^{2}EPA, Cases A+B	7.57	5.0	128.88	0.71	0.65
Case A only	7.56	4.98	128.63	0.705	0.659
Case B only	7.71	5.02	127.65	0.703	0.661

Which Loss Tail to Reject? Table 4 compares rejection of Case A, Case B, and both. Rejecting Case A alone reduces FID from 7.90 to 7.56, closely matching the result of 7.57 when both cases are rejected. Rejecting Case B alone yields a smaller improvement, with FID reaching 7.71. These results suggest that the observed improvement is mainly associated with rejecting Case A samples, consistent with the analysis in Section 3.3, which motivates filtering samples with weak alignment.

Table 5: Ablation study of R^{2}EPA on SiT-XL. (Left) Ablation study on different warmup steps. (Right) Ablation study on different rejection thresholds σ.

Warmup k_0	FID↓	sFID↓	IS↑	Prec↑	Rec↑
REPA	7.90	5.1	122.60	0.70	0.65
$k_0 = 0$	7.64	5.1	127.28	0.70	**0.66**
$k_0 = 100$K	7.67	5.0	127.86	0.71	**0.66**
$k_0 = 200$K	**7.57**	**5.0**	**128.88**	**0.71**	0.65

σ threshold	FID↓	sFID↓	IS↑	Prec↑	Rec↑
REPA	7.90	5.1	122.60	0.7	0.65
σ-reject	7.66	5.0	128.66	0.70	**0.66**
2σ-reject	7.60	5.0	128.37	0.71	**0.66**
3σ-reject	**7.57**	**5.0**	**128.88**	**0.71**	0.65

Warm-up and Threshold. Table 5 evaluates the rejection mechanism on SiT-XL with different warmup steps and rejection thresholds. All configurations improve the performance over the REPA baseline, with relatively small differences among them. The configuration with $k_0 = 200$K and $c = 3$ achieves the best FID of 7.57 among these settings. One possible explanation is that these configurations filter out similarly extreme alignment outliers, yielding comparable benefits despite differences in the rejection settings.

Qualitative Comparison. Figure 4 presents examples from REPA and R^{2}EPA with and without CFG. R^{2}EPA produces clearer object structures and more distinct local details in these examples, such as the jaguar's facial features and the golf ball's surface texture. The pumpkin and joystick also exhibit more recognisable shapes, complementing the improvements observed in the quantitative results.



Figure 4: **Qualitative Comparison.** Generated samples from REPA and R^{2}EPA on ImageNet without CFG (left) and with CFG (right). The top row shows results generated by REPA, and the bottom row shows results generated by R^{2}EPA.

Multi-Seed Results Table 6 evaluates SiT-XL across three different seeds with CFG scale 1.8 and interval guidance over $[0, 0.7]$. This combination provides a strong baseline, making further improvements particularly challenging. Even in this setting, R^{2}EPA achieves lower FID and higher IS than REPA for each evaluated seed, supporting its robustness to the choice of random seed.

Table 6: SiT-XL results across different seeds with interval CFG.

Method	Metric	Seed 1	Seed 2	Seed 3	Mean $\pm$ std.
REPA	FID↓	1.93	1.95	1.97	1.95 ± 0.02
R^{2}EPA	FID↓	1.89	1.88	1.90	1.89 ± 0.01
REPA	IS↑	281.89	281.98	281.35	281.74 ± 0.34
R^{2}EPA	IS↑	282.02	282.37	283.14	282.51 ± 0.57

6 Conclusion

We use a variational analysis to relate REPA to conditional information maximisation and examine how the alignment gradient varies with cosine similarity. This analysis motivates per-sample loss deviation as a practical indicator of atypical alignment updates. Based on this insight, we propose R^{2}EPA, a lightweight selective rejection mechanism that filters alignment-loss outliers via a statistically grounded dual-detector design with cross-batch EMA statistics. Empirical gradient measurements associate the selected samples with weaker alignment and a larger alignment-to-denoising gradient-norm ratio. R^{2}EPA requires no additional model parameters. Results on ImageNet and COCO suggest that selective alignment can benefit generation quality, while the observed training trajectories indicate potential improvements in training efficiency. Boundary tracking illustrates complementary detector roles. The Case A/B comparisons suggest a larger contribution from the Case A rejection branch in the tested setting.

REFERENCES

David Barber and Felix V. Agakov. The IM algorithm: A variational approach to information maximization. In *Advances in Neural Information Processing Systems*, volume 16, 2003.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, pp. 11472–11481. IEEE, 2022.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

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.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pp. 8780–8794, 2021.

Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953.

Hyojun Go, Yunsung Lee, Seunghyun Lee, Shinhyeok Oh, Hyeongdon Moon, and Seungtaek Choi. Addressing negative transfer in diffusion models. *NeurIPS*, 36, 2024.

Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. DINO as a von mises-fisher mixture model. In *ICLR*. OpenReview.net, 2023.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *ICCV*, pp. 7407–7417. IEEE, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

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*, pp. 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

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

Jin-Young Kim, Hyojun Go, Lea Bogensperger, Julius Erbach, Nikolai Kalischek, Federico Tombari, Konrad Schindler, and Dominik Narnhofer. Understanding, accelerating, and improving meanflow training, 2025a. URL https://arxiv.org/abs/2511.19065.

Jin-Young Kim, Hyojun Go, Soonwoo Kwon, and Hyun-Gyoon Kim. Denoising task difficulty-based curriculum for training diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=96GMFXsbJE.

Myunsoo Kim, Donghyeon Ki, Seong-Woong Shim, and Byung-Jun Lee. Adaptive non-uniform timestep sampling for diffusion model training. *CVPR*, 2025c.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, pp. 3929–3938, 2019.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *NeurIPS*, 2024.

Kyoungmin Lee, Jihun Park, Jongmin Gim, Wonhyeok Choi, Kyumin Hwang, Jaeyeul Kim, and Sunghoon Im. A training-free style-personalization via svd-based feature decomposition. *CVPR*, 2026.

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.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *ECCV*, 2024.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W. Battaglia. Generating images with sparse representations. In *ICML*, volume 139, pp. 7958–7968. PMLR, 2021.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Jihun Park, Kyoungmin Lee, Jongmin Gim, Hyeonseo Jo, Minseok Oh, Wonhyeok Choi, Kyumin Hwang, Jaeyeul Kim, Minwoo Choi, and Sunghoon Im. Infinite-story: A training-free consistent text-to-image generation. *AAAI*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Friedrich Pukelsheim. The three sigma rule. *The American Statistician*, 48(2):88–91, 1994.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685. IEEE, 2022.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure?, 2025.

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*. OpenReview.net, 2021.

Kai Wang, Yukun Zhou, Mingjia Shi, Zhihang Yuan, Yuzhang Shang, Xiaojiang Peng, Hanwang Zhang, and Yang You. A closer look at time steps is worthy of triple speed-up for diffusion model training. In *CVPR*, 2025a.

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. In *NeurIPS*, 2025b.

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, 2025.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 2026.

Hu Yu, Li Shen, Jie Huang, Hongsheng Li, and Feng Zhao. Unmasking bias in diffusion model training. In *The 18th European Conference on Computer Vision ECCV 2024*. Springer, 2024.

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.

Tianyi Zheng, Peng-Tao Jiang, Ben Wan, Hao Zhang, Jinwei Chen, Jia Wang, and Bo Li. Beta-tuned timestep diffusion model. In *ECCV*, volume 15061 of *Lecture Notes in Computer Science*, pp. 114–130. Springer, 2024.

Tianyi Zheng, Jiayang Zou, Peng-Tao Jiang, Hao Zhang, Jinwei Chen, Jia Wang, and Bo Li. Bidirectional beta-tuned diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–15, 2025. doi: 10.1109/TPAMI.2025.3604039.

APPENDIX

A THEORETICAL PROOFS

A.1 PROOF OF PROPOSITION 1

Proof. For any normalised variational conditional density $q(z_2 \mid z_1, t)$, non-negativity of conditional KL divergence gives

$$\mathbb{E}_{p(z_1,t)}[D_{\mathrm{KL}}(p(z_2 \mid z_1, t) \,\|\, q(z_2 \mid z_1, t))] \geq 0. \tag{13}$$

Expanding the conditional mutual information and making the variational gap explicit yields the Barber–Agakov bound:

$$\begin{aligned} I(Z_1; Z_2 \mid T) &= H(Z_2 \mid T) + \mathbb{E}_{p(z_1,z_2,t)}[\log q(z_2 \mid z_1, t)] \\ &\quad + \mathbb{E}_{p(z_1,t)}[D_{\mathrm{KL}}(p(z_2 \mid z_1, t) \,\|\, q(z_2 \mid z_1, t))] \\ &\geq H(Z_2 \mid T) + \mathbb{E}_{p(z_1,z_2,t)}[\log q(z_2 \mid z_1, t)]\,. \end{aligned} \tag{14}$$

We choose the vMF variational family $q(z_2 \mid z_1, t) = q_\kappa(z_2 \mid z_1)$ from equation 4. This is a valid conditional density on the unit sphere for every z_1. Substitution into equation 14 gives

$$I(Z_1; Z_2 \mid T) \geq H(Z_2 \mid T) + \log C_{d'}(\kappa) + \kappa \mathbb{E}[Z_1^\top Z_2]. \tag{15}$$

Let J be uniform on $\{1, \ldots, N\}$. By the definitions of $Z_1 = Z_1^{[J]}$, $Z_2 = Z_2^{[J]}$, and the per-sample REPA loss,

$$\mathbb{E}[Z_1^\top Z_2] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} Z_1^{[n]\top} Z_2^{[n]}\right] = -\mathbb{E}[\mathcal{L}_{\mathrm{REPA}}]. \tag{16}$$

Combining equation 15 and equation 16 proves equation 5. Because the data distribution and encoder are fixed, the distribution of Z_2 and hence $H(Z_2 \mid T)$ do not depend on (θ, ϕ). The vMF normaliser is also fixed when κ is fixed. Therefore C_κ is independent of the trainable parameters. □

A.2 PROOFS OF PROPOSITIONS 2 AND 3

Let $u_{1,i}^{[n]}$ denote the projected feature before ℓ_2 normalisation, so that $z_{1,i}^{[n]} = u_{1,i}^{[n]} / \left\| u_{1,i}^{[n]} \right\|$. The derivative of the negative cosine score with respect to this pre-normalised feature is

$$\begin{aligned} \nabla_{u_{1,i}^{[n]}}\left(-z_{1,i}^{[n]\top} z_{2,i}^{[n]}\right) &= -\frac{1}{\left\| u_{1,i}^{[n]} \right\|}\left(I - z_{1,i}^{[n]} z_{1,i}^{[n]\top}\right) z_{2,i}^{[n]} \\ &= -\frac{\delta_i^{[n]}}{\left\| u_{1,i}^{[n]} \right\|}. \end{aligned} \tag{17}$$

The parameter gradient is obtained by propagating this residual through the sample-dependent network Jacobian. Consequently, the following results establish feature-space limiting behaviour; translating it into a parameter-space statement requires the stated regularity conditions.

Proof of Proposition 2.

Proof. As $\mathrm{sim}_i^{[n]} \to 0$, the subtracted term satisfies

$$\left\| \mathrm{sim}_i^{[n]} \cdot z_{1,i}^{[n]} \right\| = \left\| \mathrm{sim}_i^{[n]} \right\| \cdot \left\| z_{1,i}^{[n]} \right\| = \left\| \mathrm{sim}_i^{[n]} \right\| \;\to\; 0, \tag{18}$$

where we used $\left\| z_{1,i}^{[n]} \right\| = 1$. Therefore $\delta_i^{[n]} \to z_{2,i}^{[n]}$ and $\left\| \delta_i^{[n]} \right\| \to 1$. Thus the feature-space residual remains at full scale. Equation 17 shows that its parameter-space image is governed by the network Jacobian. Since both the Jacobian and encoder feature vary across samples, the propagated gradient may have a negative projection onto the aggregate gradient, motivating the empirical outlier analysis. □

Proof of Proposition 3.

Proof. Expand $\left\|\delta_i^{[n]}\right\|^2$ using $\|z_1\| = \|z_2\| = 1$ and $z_{2,i}^{[n]\top} z_{1,i}^{[n]} = \mathrm{sim}_i^{[n]}$:

$$\begin{aligned}\left\|\delta_i^{[n]}\right\|^2 &= \left\|z_{2,i}^{[n]}\right\|^2 - 2\mathrm{sim}_i^{[n]}\left(z_{2,i}^{[n]\top} z_{1,i}^{[n]}\right) \\ &\quad + \left(\mathrm{sim}_i^{[n]}\right)^2 \left\|z_{1,i}^{[n]}\right\|^2 \\ &= 1 - \left(\mathrm{sim}_i^{[n]}\right)^2.\end{aligned} \tag{19}$$

Factor: $1 - (\mathrm{sim})^2 = (1 - \mathrm{sim})(1 + \mathrm{sim})$. As $\mathrm{sim}_i^{[n]} \to 1$, the first factor $1 - \mathrm{sim}_i^{[n]} \to 0$ while the second remains bounded ($1 + \mathrm{sim}_i^{[n]} \to 2$), so $\left\|\delta_i^{[n]}\right\|^2 \to 0$. If $\left\|u_{1,i}^{[n]}\right\|$ is bounded away from zero and the relevant network Jacobians are bounded, Equation 17 and the chain rule imply $\|g_i\| \to 0$. □

Note that the residual also vanishes as $\mathrm{sim}_i^{[n]} \to -1$. Therefore, although the REPA loss is maximised at $\mathrm{sim}_i^{[n]} = -1$, the residual magnitude is maximised at $\mathrm{sim}_i^{[n]} = 0$.

Vysochanskij–Petunin inequality.

Lemma 1 (Vysochanskij–Petunin (Pukelsheim, 1994), 1980)**.** *For any unimodal distribution with mean μ and standard deviation σ, and $c \geq \sqrt{8/3}$,*

$$\mathbb{P}\big(|X - \mu| > c\sigma\big) \leq \frac{4}{9c^2}. \tag{20}$$

Remark 1. As illustrated in Figure 2, the observed per-sample REPA loss distributions are approximately unimodal after warm-up. This empirical shape motivates using Lemma 1 to calibrate the rejection threshold.

B More Analysis and Results

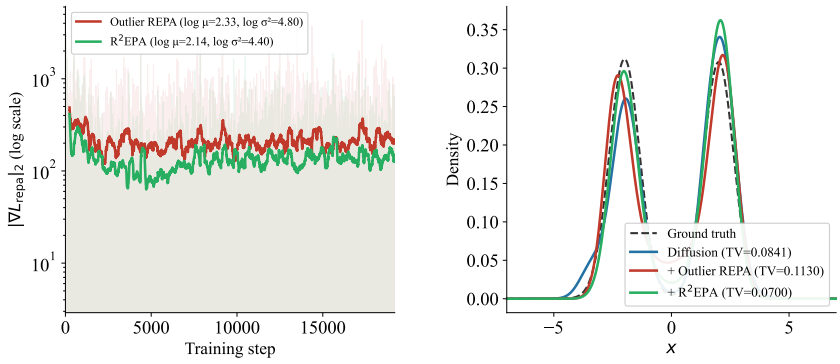

Figure 5: **Toy Experiments.** We compare three conditions: **(i)** Vanilla Diffusion, using $\mathcal{L}_{\text{diff}}$ only; **(ii)** Outlier REPA, where 2% of alignment targets are corrupted by random noise; and **(iii)** +R^{2}EPA, where our rejection rule excludes detected outliers from the alignment objective.

Toy Experiment. We construct a controlled 1D experiment to illustrate the core motivation of R^{2}EPA: that even a small fraction of harmful alignment samples can introduce large gradient instability and degrade generation quality. The training distribution is a two-component Gaussian mixture, fitted by a denoising MLP under the vanilla DDPM objective (Ho et al., 2020). We augment the diffusion loss with a feature-alignment regulariser $\mathcal{L}_{\text{repa}} = \|\hat{x}_0 - z\|^2$, where $\hat{x}_0$ is the denoiser's predicted clean sample and z is the alignment target (Yu et al., 2025). As shown in Figure 5, even the slight 2% outlier targets inflate both the mean and variance of the REPA gradient norm. This gradient

instability translates directly into degraded generation: The $\mathbb{TV}$ distance to the ground truth increases from 0.084 (vanilla diffusion) to 0.113 with outlier REPA (34% relative degradation), while R^{2}EPA, by excluding the corrupted targets, reduces $\mathbb{TV}$ to 0.070, improving over the vanilla diffusion baseline. In this controlled setting, the results illustrate the sensitivity of alignment training to corrupted targets and motivate examining sample selection in larger models.

Training Trajectories. Figure 6 compares FID for SiT-L with CFG across different training steps. R^{2}EPA achieves lower FID than REPA. At 350K steps, it reaches FID 2.76, improving upon REPA's 2.81 at 400K steps with 12.5% fewer training steps. Its FID further improves to 2.74 at 400K steps. These results suggest that selective rejection can improve training efficiency.

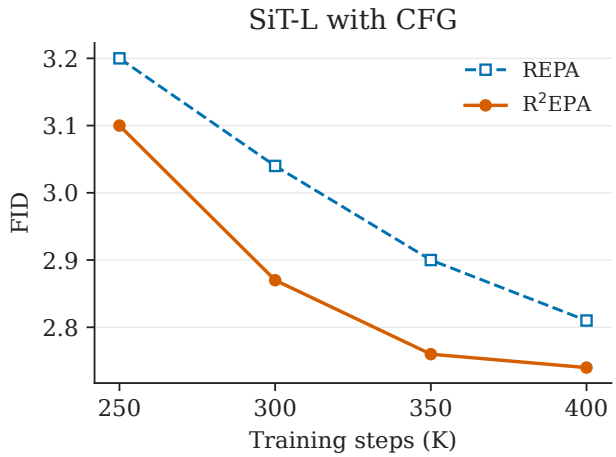

Figure 6: FID during training for SiT-L with CFG on ImageNet. R^{2}EPA achieves lower FID than REPA.

Timestep Distribution of Rejected Samples. Figure 7 shows the timestep distribution of rejected samples. The distribution is concentrated at large timesteps, with a mean of $t = 0.825$, indicating that the rejection mechanism predominantly selects samples with high noise levels. As $t \to 1$, the noisy input x_t approaches Gaussian noise, so the DiT features contain less semantic information that can be matched to the frozen encoder target. A complementary offline analysis of a standard REPA model (Appendix B) also finds that samples selected for rejection concentrate at high noise and have lower, but positive, average cosine similarity. The detector uses only per-sample loss deviation and has no explicit access to t; its concentration on high-noise timesteps therefore provides additional empirical evidence that the score captures atypical alignment conditions.

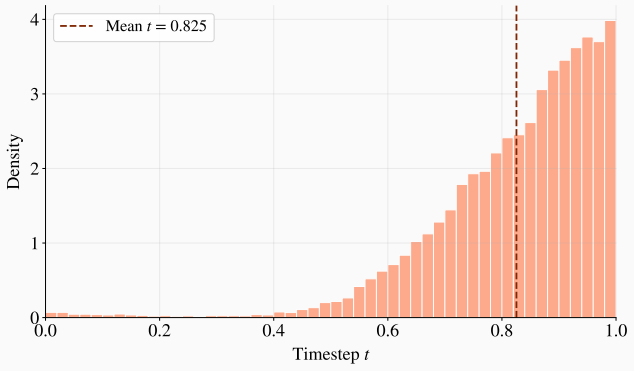

Figure 7: Distribution of timesteps for rejected samples. The rejected samples are concentrated at high noise levels, with a mean of $t = 0.825$.

Evaluation Across Diffusion Spaces. We additionally evaluate R^{2}EPA on JiT-B-16 (Li & He, 2025) with iREPA, which operates in pixel diffusion space. As shown in Table 7, adding R^{2}EPA changes FID from 55.26 to 54.93 and IS from 27.05 to 27.12, suggesting a modest benefit in this pixel-space configuration.

Table 7: Evaluation in pixel diffusion space. Comparison of iREPA and iREPA + R^{2}EPA on JiT-B-16.

Model	Method	IS↑	FID↓
JiT-B-16	iREPA	27.05	55.26
	iREPA + R^{2}EPA	**27.12**	**54.93**

Auxiliary Representation-Space Evidence We complement FID with the recently proposed FDr^6 metric (Yang et al., 2026), which averages normalised Fréchet distances across six feature spaces: Inception, ConvNeXt, DINOv2, MAE, SigLIP, and CLIP. This provides a broader assessment of the discrepancy between generated and real image distributions than using Inception features alone. As shown in Table 8, R^{2}EPA improves FDr^6 from 15.5815 to 15.5220 compared with REPA. This improvement provides complementary evidence for our method under an evaluation that incorporates multiple feature spaces, beyond the standard Inception FID.

Table 8: Auxiliary FDr^6 evaluation. Lower is better.

Metric	REPA	R^{2}EPA
FDr^6	15.5815	15.5220

Boundary Tracking Analysis Figure 3 illustrates how Z-score and MAD jointly determine the rejection boundaries. During active rejection, MAD sets the upper boundary at 64.7% of logged steps, while Z-score sets it at 35.3%. For the lower boundary, the corresponding shares are 39.1% and 60.9%. Thus, the two detectors play complementary roles, with each determining the effective boundary at different steps. Across all logged batches, samples with high alignment losses account for 99.38% of flags, compared with 0.62% for samples with low alignment losses.

Gradient Analysis. We analyse a REPA SiT-XL/2 model trained for 400K steps without sample rejection, using DINOv2 patch features as alignment targets. After training, we apply our detector without updating the model parameters. It flags 200 of 2,432 samples, of which 196 have $t > 0.5$. We measure gradients with respect to the eighth Transformer block, whose features are used for alignment. For its parameters θ_l, we define $g_i^{\text{REPA}} = \nabla_{\theta_l} \ell_i$ and $g_i^{\text{denoise}} = \nabla_{\theta_l} \ell_i^{\text{velocity}}$. The gradients are computed before applying the global REPA coefficient, and their norm ratio is averaged over individual samples within each group. Table 9 compares accepted and rejected samples with $t > 0.5$. Rejected samples have lower mean alignment similarity (0.423 vs. 0.589), together with a larger mean ratio of the alignment gradient norm to the denoising gradient norm (3.82 vs. 2.80). These observations show that the detector selects samples with comparatively weaker alignment and a larger relative alignment contribution, supporting selective control of their alignment updates.

Table 9: Gradient measurements on a REPA model for samples with $t > 0.5$. Accepted and rejected groups are determined by the detector. Values are group means except sample counts.

Quantity	Accepted	Rejected
Number of samples	1,040	196
Mean patch cosine similarity	0.589	0.423
$\|g_i^{\text{REPA}}\| / \|g_i^{\text{denoise}}\|$	2.80	3.82
$\|g_i^{\text{denoise}}\|$	0.127	0.087

C Detailed Experimental Setup

Table 10 summarises the hyperparameter settings across model scales. For the main ImageNet evaluation, we generate 50K samples to compute the evaluation metrics.

Table 10: Hyperparameter settings across different model scales.

Backbone	SiT-B	SiT-L	SiT-XL
Architecture			
#Params	132M	460M	677M
Input	$32 \times 32 \times 4$	$32 \times 32 \times 4$	$32 \times 32 \times 4$
Layers	12	24	28
Hidden dim.	768	1,024	1,152
Num. heads	12	16	16
REPA settings			
λ	0.5	0.5	0.5
Alignment depth	4	8	8
$\text{sim}(\cdot, \cdot)$	cos. sim.	cos. sim.	cos. sim.
Encoder $\mathcal{E}_{VF}(I)$	DINOv2-B	DINOv2-B	DINOv2-B
Optimization			
Batch size	256	256	256
Optimizer	AdamW	AdamW	AdamW
lr	0.0001	0.0001	0.0001
(β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Interpolants			
α_t	$1 - t$	$1 - t$	$1 - t$
σ_t	t	t	t
w_t	σ_t	σ_t	σ_t
Training objective	v-prediction	v-prediction	v-prediction
Sampler	Euler-Maruyama	Euler-Maruyama	Euler-Maruyama
Sampling steps	250	250	250

D AI Use Statement

We used generative AI tools to assist with manuscript drafting, language polishing, grammar correction, and the analysis, interpretation, and presentation of experimental results. The authors reviewed and verified the content and suggestions provided by these tools before incorporating them into the manuscript. We take full responsibility for the final content of this work.

E MORE VISUALIZATION RESULTS

In this section, we provide additional generation results without CFG in Figure 8 and generation results with CFG in Figure 9. We employ the 250-step SDE sampler in the inference stage.

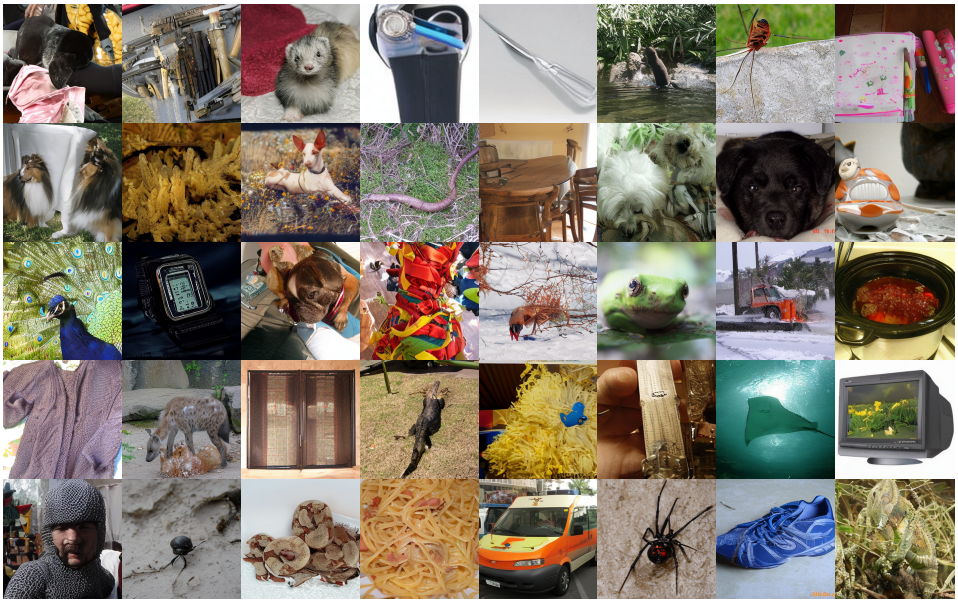

Figure 8: **Qualitative Comparison.** The generation results from R^{2}EPA (SiT-XL) on the ImageNet dataset without CFG.

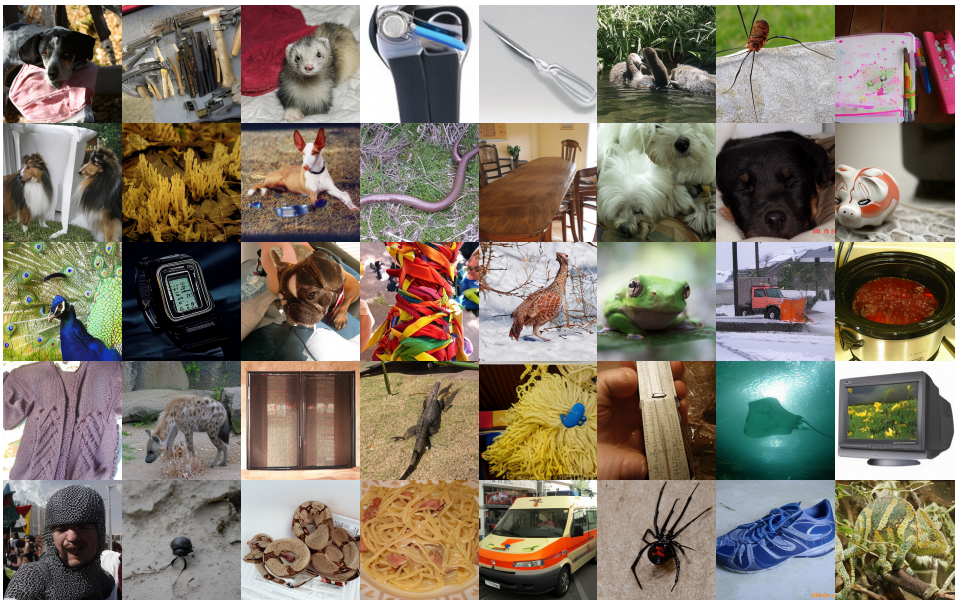

Figure 9: **Qualitative Comparison.** The generation results from R^{2}EPA (SiT-XL) on the ImageNet dataset with CFG.