FUSEREG: REGULARIZING LAYER FUSION MITIGATES THE RECONSTRUCTION-GENERATION GAP IN REPRESENTATION AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation autoencoders (RAEs) reuse features from a pretrained visual encoder as reconstruction and diffusion latents, integrating strong visual representations into image generation. However, RAEs still need to decide which encoder layers form the shared latent space for the generator and pixel decoder. This choice involves a trade-off. Shallower layers tend to preserve fine pixel details better, while deeper layers tend to yield better generation metrics. A fixed heuristic layer fusion therefore couples two stages that benefit from different information. We introduce FuseReg, which replaces heuristic feature selection with training over random subsets of encoder layers. We theoretically analyze the underlying mechanism: subset sampling explicitly penalizes sensitivity to cross-layer disagreement. On ImageNet-256 with DINOv3-L, a single FuseReg decoder reconstructs from full, sparse, and single-layer fusions without retraining, achieving higher PSNR than decoders specialized to fixed fusions. This flexibility also benefits generation: decoder replacement alone reduces unguided gFID by 27% with an unchanged RAEv2 DiT-XL generator. Applying FuseReg to both stages also reduces unguided gFID by 29% on DiT-Base. The reconstruction and generation benefits also extend to other encoder families. FuseReg narrows the reconstruction–generation gap without additional training cost or architectural changes.

1 INTRODUCTION

Representation autoencoders (RAEs) use a frozen visual encoder to define the latent space of a diffusion model and learn a decoder that maps the encoder's features back to pixels (Zheng et al., 2025). A visual encoder, however, produces a hierarchy of features whose content changes with depth, leaving the RAE to decide how those layers become a single latent. Layer fusion defines the interface between the generator, which models the latent distribution, and the decoder, which turns generated latents into images. A decoder tied to one fusion can therefore limit the quality and flexibility of the entire pipeline.

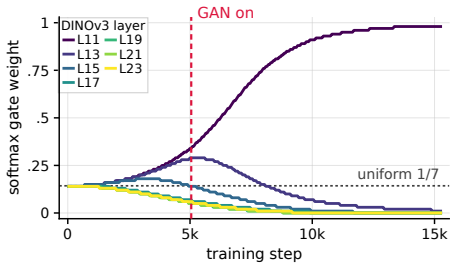


Figure 1: **A learnable gate collapses onto a shallow layer.** Per-layer gate weights over training for a softmax gate on DINOv3 layers. The shallowest layer grows to dominate while deeper layers decay to 0; the collapse accelerates once the GAN is on.

RAEv2 commits to a fixed heuristic layer fusion by summing the last k encoder layers (Singh et al., 2026). In the DINOv3-L setting (Siméoni et al., 2025), expanding the fused subset from $k{=}7$ to $k{=}23$ increases reconstruction PSNR from 22.58 to 27.04 dB but also increases unguided gFID from 1.65 to 3.01 and guided gFID from 1.06 to 1.25. The fusion that best preserves pixel quality is therefore not the one that is easiest to model generatively. Adding shallow, pixel-aligned layers supplies details that directly help reconstruction losses such as ℓ_2 and LPIPS (Zhang et al., 2018); the subset restricted to deeper, more semantic layers yields lower reported gFID in this setting. A single fixed subset binds the decoder and generator to the same choice even though they solve different

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.

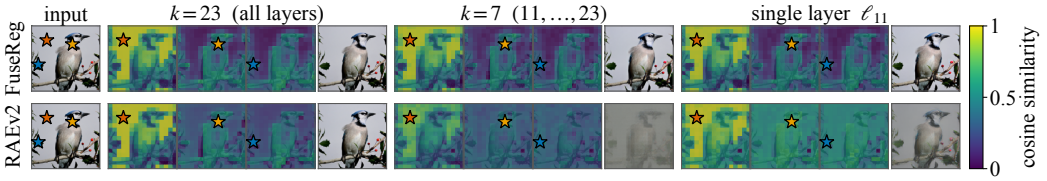


Figure 2: **Cosine-similarity maps of decoder intermediate-block features under different layer fusions.** For three query patches (stars), we visualize patch-wise cosine similarity at a decoder intermediate block, for FuseReg (top) and RAEv2 (bottom), fed three fusions: $K{=}23$, $K{=}7$, and the single layer ℓ_{11}; the rightmost column is the reconstruction. FuseReg keeps clean, stable spatial structure across all fusions, whereas RAEv2 degrades away from its $K{=}23$ training fusion.

prediction problems. The reconstruction objective favors the shallowest layers and ignores deep layers (Figure 1), whereas diffusion favors the deepest layers. This difference in latent preference is the reconstruction–generation gap that we address.

We address this mismatch with **FuseReg** (layer-**fus**ion **reg**ularization), which trains the decoder and generator on varying combinations of encoder layers. During training, we randomly select a subset of layers to form the latent. Learning to reconstruct the pixels from different layer combinations encourages the decoder to use information across layers and reduces its reliance on shallow-layer shortcuts. The resulting decoder supports any full, sparse, and single-layer fusions without retraining for each configuration (Figure 2). We extend the same principle to the generator: the DiT receives noisy subset fusions and learns to predict the full-layer representation. By regularizing layer reliance at both stages, FuseReg mitigates the reconstruction–generation gap without additional computational cost or architectural changes.

Contributions.

- **Layer-fusion regularization.** We introduce FuseReg, a layer-fusion regularizer that mitigates the reconstruction–generation gap by training the decoder and generator on normalized fusions of randomly sampled encoder layers.
- **A theoretical basis for randomized fusion.** We prove that FuseReg preserves the full-layer mean while turning cross-layer disagreement into an explicit regularization penalty. We further establish a second-order separation from any deterministic global fusion, including heuristic layer selection and learnable global gates.
- **Robust reconstruction and improved generation.** On ImageNet-256 with DINOv3-L, one FuseReg decoder reconstructs from full, sparse, and single-layer fusions. Decoder replacement alone reduces unguided gFID from 3.01 to 2.21 at $k{=}23$, while joint regularization reduces DiT-Base gFID from 13.96 to 9.93. These gains also extend to other pretrained encoder families.

2 Related Work

Existing representation-based generative methods primarily improve performance by refining the latent representation or its alignment with the generative model. RAEs use frozen visual encoders to define the latent space of generative models (Zheng et al., 2025), and subsequent work extends this paradigm to video generation (Guo et al., 2026; Xie et al., 2026). Building on this framework, one line of work redesigns the tokenizer or latent space (Jia et al., 2026; Bi et al., 2026; Chen et al., 2026; Chang et al., 2026), while another modifies the training objectives of the tokenizer or generator to improve representation alignment (Yu et al., 2025; Leng et al., 2025; Singh et al., 2025; Shin et al., 2026). These approaches improve the quality or modelability of the latent representation, but generally optimize around a single, predetermined representation configuration.

More directly related to our work are methods that aggregate features from multiple encoder depths. RAEv2 uses a fixed sum of the last several encoder layers (Singh et al., 2026), whereas DRoRAE and attentive multi-layer probing learn deterministic fusion weights across depths (Zhu et al., 2026b; Ciernik et al., 2026). These methods share the objective of selecting or learning a preferred layer

composition. In contrast, we ask how the decoder and generator can remain effective across a family of layer-fusion configurations, rather than how to identify a single optimal fusion.

From a training perspective, FuseReg is related to dropout and nested dropout (Srivastava et al., 2014; Rippel et al., 2014), as well as LayerDrop and Matryoshka representation learning (Fan et al., 2020; Kusupati et al., 2022). These methods stochastically vary the available computational paths or representation subsets during training to support regularization, flexible deployment, or efficient inference. FuseReg applies this principle to the layer-fusion interface of an RAE: we randomize the frozen encoder layers used to construct the latent representation, training downstream models for robustness across fusion configurations rather than dependence on a single fixed fusion. §A provides paper-level comparisons with representative methods.

3 FuseReg: Randomized Layer Fusion as Regularization

FuseReg targets dependence on a fixed fusion rather than searching for a different fixed fusion. We first place existing fixed fusions in a common formulation, then construct a mean-preserving distribution over layer subsets, and finally apply it at the two prediction stages of an RAE.

3.1 From a fixed fusion to a distribution over layer subsets

A fixed fusion exposes each downstream model to only one layer composition during training. We express fixed subsets and learned global gates using common fusion weights to highlight their shared limitation in learning robustness to changes in layer composition.

Given an image x and a frozen encoder E, let $h_k = \mathrm{LN}(E_{l_k}(x)) \in \mathbb{R}^{N\times d}$ be the layer-normalized tokens for each of K chosen layers ($l_k \in \{l_1, \ldots, l_K\}$), with N patch tokens and feature dimension d. A layer-fusion rule maps these features to the single latent consumed by the pixel decoder or diffusion transformer. We write a normalized fusion as

$$z_a = \sum_{k=1}^{K} a_k h_k, \qquad \mathbf{1}^\top a = 1, \tag{1}$$

where the weights a are chosen independently of the input. We use normalized weights throughout the analysis. A fixed prefix sum selects the same subset but uses a different global scale; empirical baselines retain their stated scale convention.

Fixed single-layer and prefix fusions use a fixed weight vector a; a learned global gate optimizes these weights but shares them across inputs. In our experiments, the learned gate concentrates on a single-layer shortcut (Figure 1). FuseReg instead samples layer subsets during training, exposing the downstream consumer to different layer compositions.

3.2 Training on normalized random-subset means

FuseReg constructs a distribution over layer subsets that varies the layer composition during training while preserving the deployment latent in expectation. We sample a separate layer-dropout mask for each training example. For $p \in [0, 1)$, draw a raw i.i.d. Bernoulli mask $\widetilde{m}$ and condition it on being nonzero:

$$\begin{aligned} \widetilde{m}_k &\overset{\text{iid}}{\sim} \mathrm{Bernoulli}(1-p), \qquad k = 1, \ldots, K, \\ m \sim \mathcal{M}_p &:= \mathrm{Law}\big(\widetilde{m} \,\big|\, \mathbf{1}^\top \widetilde{m} > 0\big), \qquad z_m = \frac{\sum_k m_k h_k}{\sum_k m_k}. \end{aligned} \tag{2}$$

The corresponding fusion weights are $a_k = m_k / \sum_j m_j$. For $0 < p < 1$, the sampling distribution has support on all $2^K - 1$ nonempty subsets. At inference, all layers are retained by default and z_m becomes the full-layer mean $\bar{h} = K^{-1} \sum_k h_k$; subsets are used to test robustness.

The normalization in Equation (2) is essential. By linearity of expectation, $\mathbb{E}_m[z_m \mid x] = \bar{h}$: changing p varies the disagreement around the deployment latent without shifting that latent in expectation. In §4, we prove that this stochastic operation is exactly a layer-disagreement regularizer for a linear squared-error consumer; complete derivations are provided in §F.

3.3 TWO PREDICTION STAGES, TWO REGULARIZATION RATES

The drop rate p controls the mean-preserving distribution over layer fusions. We apply FuseReg to the decoder, which maps a fused latent back to pixels, and the diffusion transformer (DiT), which models and generates latents. Because the two stages solve different prediction problems, we do not assume that the same drop rate is optimal for both. We introduce separate drop rates, p_{dec} and p_{dit}. The two stages share the regularization mechanism, but its strength can be set separately for their different prediction targets.

Decoder: subset fusions to pixels (p_{dec}). FuseReg trains the pixel decoder to reconstruct images from multiple layer compositions. A ViT decoder D reconstructs $\hat{x} = D(z_m)$ with $m \sim \mathcal{M}_{p_{\mathrm{dec}}}$, under a pixel, perceptual, and adversarial objective (Zhang et al., 2018; Goodfellow et al., 2014): $\mathcal{L}_{\mathrm{dec}} = \|\hat{x} - x\|_2 + \lambda_{\mathrm{p}}\mathcal{L}_{\mathrm{LPIPS}}(\hat{x}, x) + \lambda_{\mathrm{g}}\mathcal{L}_{\mathrm{GAN}}(\hat{x})$.

DiT: subset latents to the full-layer fusion (p_{dit}). FuseReg varies the layer composition entering the noisy flow path while retaining one common prediction target. The generative stage uses the same masking construction, but its target is the full-layer aggregate $z_{\mathrm{full}} = K^{-1}\sum_k h_k$. With a flow-matching interpolant (Lipman et al., 2023) $z_t = tz_m + (1-t)\epsilon$ built from a subset aggregate ($m \sim \mathcal{M}_{p_{\mathrm{dit}}}$, $t \sim \nu$, $\epsilon \sim \mathcal{N}(0, I)$, class label c), the DiT regresses the full-layer target with the time-weighted x-prediction loss:

$$\mathcal{L}_{\mathrm{DiT}} = \mathbb{E}_{(x,c),m,t,\epsilon}\rho(t)\big\|x_\theta(z_t, t, c) - z_{\mathrm{full}}\big\|_2^2. \tag{3}$$

Here $\rho(t) = \max(1-t, 0.05)^{-2}$ and ν is the shifted logit-normal law, expressed in the noise-to-data time direction. Both output heads use this loss; §F.2 gives the time law and its relation to the RAEv2 velocity-space loss (Singh et al., 2026). Setting $p_{\mathrm{dit}} = 0$ recovers the standard full-fusion objective; any $p_{\mathrm{dit}} > 0$ trains the generator to map multiple subset fusions to the same full-layer target.

The decoder and DiT therefore see the same type of layer perturbation inside different prediction problems. The pair $(p_{\mathrm{dec}}, p_{\mathrm{dit}}) \in [0, 1)^2$ is the exchangeable two-dimensional slice of the per-layer rate relaxation described in §F.5.

4 WHY FUSEREG TARGETS LAYER DISAGREEMENT

Random subset fusion is useful only if its variation targets the brittle part of the decoder–generator interface rather than arbitrarily corrupting the fused latent. We show that normalized subset sampling leaves the all-layer latent unchanged in expectation and adds variation precisely along directions on which encoder layers disagree. We use a linear squared-loss model as an analytically tractable surrogate to study how this variance regularizes pixel decoding and latent prediction differently. Formal statements are presented below, with complete proofs and extended analysis in §F.

4.1 WHAT SUBSET SAMPLING CHANGES

We first isolate the statistical effect of FuseReg independently of any decoder or generator. Vectorize each layer feature h_k and define the all-layer mean, layer deviations, and their covariance by

$$\bar{h} = \frac{1}{K}\sum_{k=1}^{K} h_k, \qquad e_k = h_k - \bar{h}, \qquad V(x) = \frac{1}{K}\sum_{k=1}^{K} e_k e_k^\top, \qquad \Gamma = \mathbb{E}_x[V(x)]. \tag{4}$$

Let S be the nonempty subset sampled by Equation (2) with cardinality $R = |S|$. We define the subset mean latent representation as $z_S = R^{-1}\sum_{k\in S} h_k$. The expected sampling variance coefficient $c_K(p)$ is then defined by

$$c_K(p) = \mathbb{E}\left[\frac{K-R}{R(K-1)}\right], \tag{5}$$

where the expectation is under a $\mathrm{Binomial}(K, 1-p)$ law conditioned on $R > 0$.

Lemma 1 Normalized subset fusion preserves the mean and isolates layer disagreement

For every fixed input x and $K \geq 2$,

$$\mathbb{E}_S[z_S \mid x] = \bar{h}, \qquad \mathbb{E}_S\big[(z_S - \bar{h})(z_S - \bar{h})^\top \mid x\big] = c_K(p)V(x). \tag{6}$$

Consequently, the sampling-induced covariance averaged over inputs is $c_K(p)\Gamma$.

Insight. Random subsets preserve the all-layer representation on average. They vary only in components where layers disagree, leaving shared components unchanged. The drop rate controls this variation, giving us a way to perturb layer disagreement without shifting the average representation.

A detailed proof is given in §F.1. This sampling identity is exact and does not assume a linear downstream model.

4.2 Why the two stages need different rates

We next place the same perturbation inside the decoder and DiT objectives. The following result is exact for homogeneous linear predictors under squared loss; its role is to identify the training signal induced by subset sampling, not to replace the nonlinear experiments.

Proposition 1 Subset fusion induces stage-specific disagreement penalties

Let y be a target with finite second moment, let D and W be homogeneous linear maps, and let $\epsilon \sim \mathcal{N}(0, I)$ be independent of (x, S).

Decoder. For squared reconstruction error,

$$\begin{aligned}\mathcal{L}_p^{\mathrm{dec}}(D) &:= \mathbb{E}_{x,S}\|Dz_S - y\|^2 \\ &= \underbrace{\mathbb{E}_x\|D\bar{h} - y\|^2}_{\mathcal{L}_0^{\mathrm{dec}}(D)} + c_K(p)\underbrace{\operatorname{tr}(D\Gamma D^\top)}_{\Omega(D)},\end{aligned} \tag{7}$$

where $\Omega(D) = K^{-1}\sum_k \mathbb{E}_x\|D(h_k - \bar{h})\|^2$.

DiT. At a fixed $t \in (0, 1]$, define $z_{t,S} = tz_S + (1-t)\epsilon$ and $\bar{z}_t = t\bar{h} + (1-t)\epsilon$. With the weight $\rho(t)$ in Equation (3), prediction of the all-layer target $\bar{h}$ gives

$$\begin{aligned}\mathcal{L}_{p,t}^{\mathrm{dit}}(W) &:= \mathbb{E}_{x,S,\epsilon}\rho(t)\|Wz_{t,S} - \bar{h}\|^2 \\ &= \underbrace{\rho(t)\mathbb{E}_{x,\epsilon}\|W\bar{z}_t - \bar{h}\|^2}_{\mathcal{L}_{0,t}^{\mathrm{dit}}(W)} + \rho(t)t^2 c_K(p)\operatorname{tr}(W\Gamma W^\top).\end{aligned} \tag{8}$$

Insight. In the linear model, subset sampling adds a cost for predictions that change with layer disagreement. The decoder reconstructs pixels, while the DiT predicts the full-layer latent from a noisy input. Separate drop rates let us tune this cost for each task.

A detailed proof is given in §F.2; §5 evaluates the effects in the nonlinear models.

4.3 Why a fixed fusion is not equivalent

Finally, we compare fusion rules through the first two moments of their layer weights. FuseReg preserves the mean while varying the weights in every layer-contrast direction. No deterministic global fusion can match these weight moments.

Let $w_S = \mathbf{1}_S/|S|$ be the random fusion weights, where $\mathbf{1}_S$ is the binary indicator vector of S, and let $P = I - K^{-1}\mathbf{1}\mathbf{1}^\top$ be the projector onto the layer-contrast subspace.

Proposition 2 No deterministic global fusion matches random-subset weight moments

For $K \geq 2$ and $0 < p < 1$,

$$\mathbb{E}[w_S] = \frac{1}{K}\mathbf{1}, \qquad \mathbb{E}[w_S w_S^\top] = \frac{1}{K^2}\mathbf{1}\mathbf{1}^\top + \frac{c_K(p)}{K}P. \tag{9}$$

The second moment in Equation (9) has rank K, while every deterministic input-independent fusion has second moment $ww^\top$ of rank at most 1. Hence no fixed global fusion matches both moments of normalized subset sampling.

Insight. Random subsets match the average weights of full fusion while varying each layer's contribution. Fixed weights cannot match both properties. This identifies what randomization adds beyond selecting or learning a single global fusion.

A detailed proof is given in §F.4. Distinct weight moments induce different linear objectives when feature moments preserve the distinction (Lemma 2). The comparison concerns input-independent fusion rules. For $0 < p < 1$, every nonempty subset receives positive probability; per-layer rates recover deterministic subsets at the vertices of their rate space (§F.5).

5 Experiments

Experimental setup. We study FuseReg through three experiments. We evaluate one decoder across layer fusions, swap decoders while fixing the DiT and sampled latents, and regularize both stages jointly. Our primary experiments use a frozen DINOv3-L (Siméoni et al., 2025) with $K{=}23$ candidate transformer layers and train a ViT decoder from scratch on ImageNet-256 (Russakovsky et al., 2015). All decoder variants are matched in architecture, parameter count, optimization hyperparameters, and training-step budget; the DiT sweeps enforce the same controls within each of the DiT-Base and DiT-XL scales. We report PSNR, SSIM (Wang et al., 2004), and reconstruction FID (rFID) (Heusel et al., 2018) on 50k images, and generation FID (gFID) and Inception Score (IS) (Salimans et al., 2016) on 50k samples from a class-conditional DiT (Peebles & Xie, 2023) with 50 Euler steps. Full protocols and baseline provenance are given in §B.

5.1 Reconstruction robustness across subsets

We begin with the prerequisite for a robust decoder–generator interface: one decoder must convert plausible full, subset, and single-layer fusions into images without retraining. This experiment isolates that capability before introducing generated latents.

Table 1: **One FuseReg decoder matches or outperforms any deterministic fusion.** Light-blue cells mark the fusion(s) seen during training: each RAEv2 decoder is trained on one deterministic aggregate ($k{=}7$ or $k{=}23$). The FuseReg decoder is trained across all nontrivial fusions.

	fusion $k{=}7$			fusion $k{=}23$			fusion ℓ_{11}		
Decoder	PSNR↑	SSIM↑	rFID↓	PSNR↑	SSIM↑	rFID↓	PSNR↑	SSIM↑	rFID↓
RAEv2$_{K=7}$	22.58	0.626	**0.30**	18.37	0.531	1.62	17.50	0.491	6.35
RAEv2$_{K=23}$	12.51	0.369	16.10	27.04	0.806	**0.18**	14.31	0.447	11.21
FuseReg$_{p=.95}$	**23.77**	**0.678**	0.60	**27.52**	**0.826**	0.42	**25.13**	**0.735**	**0.45**

Table 1 evaluates every decoder on three fusions—$k{=}7$, $k{=}23$, and the single layer ℓ_{11}—with light-blue cells marking each row's training support. Each fixed-fusion RAEv2 decoder specializes to its training fusion: RAEv2$_{K=23}$ reaches 27.04 dB and 0.18 rFID on its training fusion but drops to 12.51–14.31 dB on the others; RAEv2$_{K=7}$ shows the same pattern. A single FuseReg decoder remains strong across all three (23.77/27.52/25.13 dB), with rFID below 0.6. Qualitative outputs in Figure C.1 show the same robustness. Figure 2 shows cosine-similarity maps for three query patches in an intermediate decoder block. Off-training fusions disrupt object alignment for RAEv2, whereas FuseReg preserves it across all three fusions.

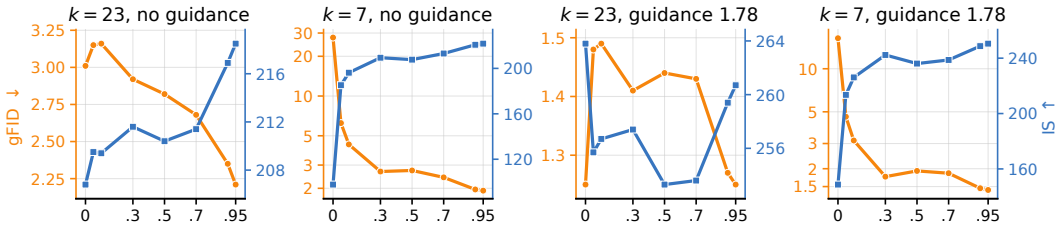


Figure 3: **Fixed generator, swapped decoders.** A single reproduced RAEv2 DiT (p_{dit}=0) produces one fixed set of latents in the k=23 or k=7 fusion space; each point renders those same latents with a decoder trained at FuseReg rate p_{dec} (horizontal axis), so changes in the rendered images arise solely from decoder replacement. Orange: gFID↓ (left axis); blue: IS↑ (right axis). The p_{dec}=0 point is the plain RAEv2 decoder. gFID/IS over 50k samples; guidance is the official RAEv2 internal guidance (scale 1.78) (Singh et al., 2026); k=7 panels use a log gFID scale. Full numbers are in Table D.1.

5.2 Generation gains from decoder replacement

To test whether reconstruction robustness improves generation, we sample latents from RAEv2 k=23 and k=7 DiTs and render them with decoders trained at increasing p_{dec}. Within each fusion and guidance setting in Figure 3, the sampled latents remain fixed and only the decoder changes.

On the native k=23 fusion, swapping the plain RAEv2 decoder for a FuseReg decoder reduces gFID from 3.01 to 2.21; under the shifted k=7 fusion, it reduces gFID from 27.73 to 1.92. The curves are not uniformly monotone. For the native k=23 fusion, unguided gFID initially worsens slightly, and guided gFID temporarily regresses from 1.25 to as high as 1.49 at intermediate rates, then returns to 1.25 at p_{dec}=0.95. Thus the strongest regularized decoder preserves the guided baseline while delivering the large unguided gain. Under shifted k=7, both unguided and guided generation improve substantially, from 27.73 to 1.92 and from 16.35 to 1.42, respectively. We examine the different guided and unguided trajectories in §6.

5.3 Generation gains from joint regularization

We test whether FuseReg improves diffusion training and whether applying it to both stages yields complementary gains. We evaluate grids over $(p_{\text{dec}}, p_{\text{dit}})$ at K=23 for both DiT-Base and DiT-XL without guidance.

Complementary gains on DiT-Base. On DiT-Base (Table 2a), regularizing either the decoder (p_{dec}=0.9, p_{dit}=0) or the generator (p_{dec}=0, p_{dit}=0.7) improves gFID from the 13.96 baseline to 12.96 and 12.09, respectively. Applying FuseReg jointly reaches 9.93 at (p_{dec}=0.9, p_{dit}=0.7). This joint improvement exceeds the additive gains of the single-axis changes (a pattern mirrored in IS, Table 2b), confirming that stage-specific drop rates act complementarily.

Scaling to DiT-XL reveals metric-dependent dynamics. When scaling to the stronger DiT-XL baseline (gFID 2.91), the generator's response to regularization becomes nuanced (Table 2c). For gFID, regularizing the generator alone fails to yield improvements; however, joint regularization with the decoder successfully unlocks further gFID gains. IS follows a different pattern. Increasing either p_{dit} or p_{dec} can improve IS.

Drop rate amplifies disagreement variance, not signal loss. Because the subset mean remains unbiased, a high drop rate like p_{dec}=0.95 does not mean "95% of the signal is destroyed." Instead, increasing p amplifies $c_K(p)$ in our linear surrogate (Equations (5) and (7)), severely penalizing layer-disagreement directions. This theoretical property manifests empirically as a non-monotonic trend along the decoder axis: at p_{dit}=0, gFID initially degrades (e.g., 17.28 at p_{dec}=0.05) before outperforming the baseline at higher rates. Notably, a sufficiently regularized generator ($p_{\text{dit}} \geq 0.5$) stabilizes this dynamic, allowing nearly all non-zero decoder rates to surpass the $(0, 0)$ baseline.

The benefits of FuseReg also extend to other pretrained encoder families (§E).

Table 2: **FuseReg across two stages and generator scales.** Unguided generation on class-conditional ImageNet-256, 40 epochs, 50k samples. Rows: generator rate p_{dit}; columns: decoder rate p_{dec}. The green box at $(0, 0)$ is the reproduced RAEv2 fixed-fusion baseline. Shading diverges from that baseline (Blue = better, Orange = worse). **Bold**: best per panel.

(a) DiT-Base, gFID ↓.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0	0.05	0.1	0.3	0.5	0.7	0.9
0	13.96	17.28	17.26	16.05	15.28	14.32	12.96
0.05	13.52	16.03	16.08	15.00	14.28	13.48	12.21
0.1	12.37	14.85	14.93	13.97	13.25	12.52	11.37
0.3	12.45	14.12	14.20	13.33	12.72	12.09	10.99
0.5	12.27	13.47	13.57	12.72	12.18	11.52	10.50
0.7	12.09	12.71	12.83	12.04	11.53	10.96	**9.93**
0.9	13.84	13.63	13.71	12.90	12.44	11.80	10.68

(b) DiT-Base, IS ↑.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0	0.05	0.1	0.3	0.5	0.7	0.9
0	107.5	100.3	100.3	103.0	105.9	107.7	111.5
0.05	108.5	102.8	102.5	105.4	108.0	109.4	113.6
0.1	114.1	108.0	108.0	110.2	113.0	114.7	118.5
0.3	113.4	109.5	109.5	111.7	114.3	115.7	119.1
0.5	114.1	112.1	112.2	114.4	116.9	118.2	121.7
0.7	116.8	116.6	116.7	118.8	121.2	122.3	**125.7**
0.9	111.4	113.8	113.7	115.9	117.8	120.1	123.5

(c) DiT-XL, gFID ↓.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0	0.05	0.1	0.3	0.5	0.7	0.9	0.95
0	2.91	3.21	3.22	3.00	2.91	2.76	2.50	**2.38**
0.5	2.91	3.19	3.21	3.02	2.93	2.80	2.56	2.45
0.7	2.96	3.13	3.17	2.97	2.90	2.77	2.53	2.42
0.9	3.09	3.19	3.24	3.01	2.94	2.76	2.50	2.39

(d) DiT-XL, IS ↑.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0	0.05	0.1	0.3	0.5	0.7	0.9	0.95
0	207.4	204.4	204.5	206.8	207.7	208.9	211.1	212.8
0.5	209.8	206.2	206.1	207.8	209.3	210.0	212.1	213.8
0.7	210.7	207.5	207.7	208.5	210.1	211.4	213.8	215.6
0.9	213.5	212.1	212.1	214.0	215.3	216.5	219.2	**221.1**

6 Discussion

How FuseReg changes layer reliance. Figures C.2 and 4 reveal a change in how the decoder accesses reconstruction information. FuseReg reconstructs effectively from individual layers and small subsets, exhibits a flatter leave-one-out sensitivity profile, and reaches diminishing marginal gains with fewer fused layers. The broadly preserved Shapley ordering (Figure C.2b) further suggests that layers differ in their value for reconstruction: a layer can remain useful across subsets while becoming less indispensable to the full fusion. Because the encoder is frozen, these changes reflect the decoder's ability to recover information already available across depths. This provides an interpretation of the decoder-swap gains (Figure 3): improving the readout of a fixed representation can improve generation even when the generator and sampled latents are unchanged.

Why the two stages respond differently. The shared disagreement covariance in Proposition 1 acts within different prediction problems: the decoder reconstructs pixels across alternative fusions, whereas the DiT predicts a full-layer target along a noisy trajectory, with a time-dependent regularization strength (Equation (8)). The rate sweeps show how this distinction matters in practice. Applying FuseReg jointly gives the best measured DiT-Base result, while DiT-XL favors strong decoder regularization and $p_{\text{dit}} = 0$ for the lowest gFID (Table 2) and $p_{\text{dit}} = 0.9$ for the highest IS. Thus the benefit of a robust decoder persists across these scales, while the preferred generator rate changes. These findings motivate choosing the two rates separately and comparing their effects over longer training schedules to clarify the roles of generator capacity and convergence.

FuseReg under internal guidance. The guided and unguided decoder-swap curves (Figure 3) suggest that the benefit of decoder regularization also depends on the sampler's latent distribution. Generator regularization introduces an additional interaction between the predictions used for guidance. RAEv2's internal guidance (Singh et al., 2026) combines the full DiT output with a prediction from an intermediate REPA head:

$$\hat{x}_{\text{guided}} = \hat{x}_{\text{full}} + w\big(\hat{x}_{\text{full}} - \hat{x}_{\text{repa}}\big).$$

The correction therefore depends on the difference between the two predictions. Changes in their relative responses to layer subsets are amplified by the guidance strength w, even when one prediction becomes individually more stable.

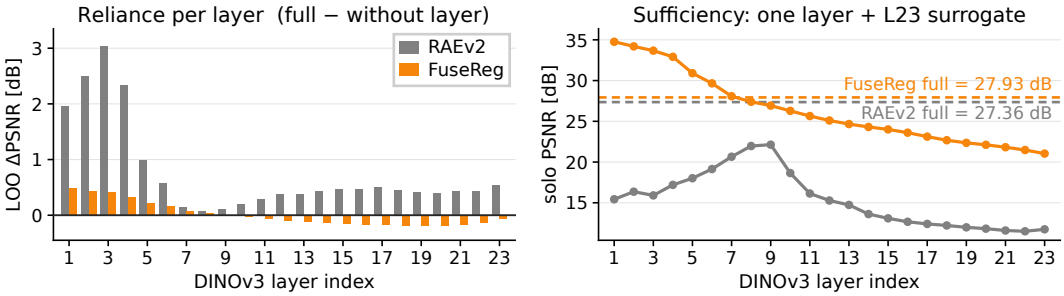


Figure 4: **FuseReg distributes reconstruction across layers.** Frozen DINOv3, K=23: the RAEv2 decoder and FuseReg (p_{dec}=0.95), evaluated on the same 10,000 held-out ImageNet-256 images with the fixed ℓ_{23} token-mean surrogate. **Left:** removing individual layers yields a flatter PSNR-drop profile for FuseReg. **Right:** individual layers yield stronger reconstructions; dashed lines mark full-fusion PSNR.

The linear surrogate in Equation (8) makes this coupling explicit. Let W_{full} and W_{repa} represent the full and REPA predictors, and write $\Omega(W) = \text{tr}(W\Gamma W^\top)$. Their guided combination is $W_{\text{guided}} = (1+w)W_{\text{full}} - wW_{\text{repa}}$, with disagreement penalty

$$\Omega(W_{\text{guided}}) = (1+w)^2\Omega(W_{\text{full}}) + w^2\Omega(W_{\text{repa}}) - 2w(1+w)\,\text{tr}\big(W_{\text{full}}\Gamma W_{\text{repa}}^\top\big).$$

The sampling-induced term is $\rho(t)t^2 c_K(p_{\text{dit}})\Omega(W_{\text{guided}})$. The cross-term measures how the two predictors respond to the same layer-disagreement directions: aligned responses partially cancel through subtraction, whereas opposing responses reinforce each other. Thus, reducing the full predictor's sensitivity alone does not determine the sensitivity of the guided combination. This provides a mechanism to investigate alongside the guidance-dependent rate preferences in Tables 2 and D.2.

7 Limitations

Our experiments focus on ImageNet-256; higher resolutions, other data domains, and longer training schedules remain to be evaluated. Preferred decoder and generator rates depend on model scale, metric, and guidance (Tables 2 and D.2), so the two rates should be selected separately for each setting. The sampling identities and second-order separation from deterministic global fusion are exact for the specified subset distribution. The objective decompositions assume homogeneous linear predictors and squared loss (§F.6). They characterize the induced training signal, while the reconstruction and generation benefits in nonlinear models are established empirically. Extending this characterization to mixed reconstruction losses and the complete guided sampling trajectory remains open.

8 Conclusion

We introduced FuseReg, which treats layer fusion as a training distribution within representation autoencoders. Sampling normalized subsets of frozen encoder features preserves the full-layer latent in expectation while varying it along cross-layer disagreement directions. This variation induces an explicit disagreement penalty under linear squared loss, and the second-order structure of subset sampling cannot be matched by deterministic global fusion. The experiments show that a single trained decoder supports full, sparse, and single-layer fusions, and that improving this readout benefits generation even with an unchanged generator. Applying FuseReg to diffusion training demonstrates complementary gains with separately chosen regularization strengths across encoder families. Together, these findings connect the reconstruction–generation gap to the dependence of downstream models on a particular layer composition. Training for robustness across compositions makes information distributed across encoder depths more accessible to both stages, providing a way to improve the decoder–generator interface while keeping the pretrained encoder fixed.

AI USE STATEMENT

The authors developed the core research ideas and the proposed method, and wrote its core implementation. We used Claude for codebase maintenance and for writing and revising experimental and supporting scripts. Both the core implementation and AI-assisted code were checked through AI-assisted review and human verification, with the authors reviewing and approving the final code.

We used GPT-6 Astra to assist with mathematical formulation and proof development, including drafting and refining derivations and checking mathematical arguments. The authors determined what needed to be proved, developed the underlying reasoning, and chose the proof strategies. All final mathematical formulations, claims, and proofs were reviewed and proofread by multiple authors and experts to check their correctness, assumptions, and clarity.

Generative AI tools also assisted with drafting and revising explanatory passages, language editing, and LaTeX preparation. The authors reviewed and edited all AI-assisted text. We did not use generative AI to create synthetic datasets or alter the experimental data.

The authors take full responsibility for the entire paper and accompanying artifacts, including the method, code, experiments, reported numerical results, tables, figures, mathematical formulations, claims and proofs, and all conclusions. All scientific judgments and final decisions remained with the authors.

REPRODUCIBILITY STATEMENT

We will publicly release all source code, model weights, training scripts, and scripts to reproduce every table in this paper. Other researchers have independently verified our table results in separate computing environments and checked the reproducibility of every reported numerical entry.

REFERENCES

Tianci Bi, Xiaoyi Zhang, Yan Lu, and Nanning Zheng. VFM-VAE: Vision foundation models can be good tokenizers for latent diffusion models, 2026. URL https://arxiv.org/abs/2510.18457.

Ramón Calvo-González and François Fleuret. Laminating representation autoencoders for efficient diffusion, 2026. URL https://arxiv.org/abs/2602.04873.

Hun Chang, Byunghee Cha, and Jong Chul Ye. Hyperspherical autoencoder for high-fidelity image reconstruction and generation, 2026. URL https://arxiv.org/abs/2601.22904.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. AlignTok: Aligning visual foundation encoders to tokenizers for diffusion models, 2026. URL https://arxiv.org/abs/2509.25162.

Laure Ciernik, Marco Morik, Lukas Thede, Luca Eyring, Shinichi Nakajima, Zeynep Akata, and Lukas Muttenthaler. Attentive multi-layer fusion for vision transformers, 2026. URL https://arxiv.org/abs/2601.09322.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020.

Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, and Ziwei Liu. The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding, 2026. URL https://arxiv.org/abs/2512.19693.

Yue Gong, Hongyu Li, Shanyuan Liu, Bo Cheng, Yuhang Ma, Liebucha Wu, Xiaoyu Wu, Manyuan Zhang, Dawei Leng, Yuhui Yin, and Lijun Zhang. RPiAE: A representation-pivoted autoencoder enhancing both image generation and editing, 2026. URL https://arxiv.org/abs/2603.19206.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

Minghui Guo, Shengqiong Wu, and Hao Fei. V-RAE: Rethinking video latent spaces for generation, 2026. URL https://arxiv.org/abs/2608.13556.

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, 2018. URL https://arxiv.org/abs/1706.08500.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, pp. 646–661. Springer, 2016.

Mingkai Jia, Mingxiao Li, Zhijian Shu, Anlin Zheng, Liaoyuan Fan, Jiaxin Guo, Tianxing Shi, Dongyue Lu, Zeming Li, Xiaoyang Guo, Xiaojuan Qi, Xiao-Xiao Long, Qian Zhang, Ping Tan, and Wei Yin. DINO-Tok: Adapting DINO for visual tokenizers, 2026. URL https://arxiv.org/abs/2511.20565.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself, 2024. URL https://arxiv.org/abs/2406.02507.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 562–570. PMLR, 2015.

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, 2025. URL https://arxiv.org/abs/2504.10483.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Siyu Liu, Chujie Qin, Hubery Yin, Qixin Yan, Zheng-Peng Duan, Chen Li, Jing Lyu, Chun-Le Guo, and Chongyi Li. Improving reconstruction of representation autoencoder, 2026. URL https://arxiv.org/abs/2602.08620.

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

Oren Rippel, Michael Gelbart, and Ryan Adams. Learning ordered representations with nested dropout. In *International Conference on Machine Learning*, pp. 1746–1754. PMLR, 2014.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs, 2016. URL https://arxiv.org/abs/1606.03498.

Jaeyo Shin, Jiwook Kim, and Hyunjung Shim. Representation alignment for just image transformers is not easier than you think, 2026. URL https://arxiv.org/abs/2603.14366.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

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. URL https://arxiv.org/abs/2512.10794.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders, 2026. URL https://arxiv.org/abs/2605.18324.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Zhihao Xie, Junfeng Wu, Xinting Hu, Junchao Huang, and Li Jiang. VideoRAE: Taming video foundation models for generative modeling via representation autoencoders, 2026. URL https://arxiv.org/abs/2607.14088.

Jiahui Yu and Thomas S. Huang. Universally slimmable networks and improved training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1803–1811, 2019.

Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. In *International Conference on Learning Representations*, 2019.

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, 2025. URL https://arxiv.org/abs/2410.06940.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

Shilong Zhang, He Zhang, Zhifei Zhang, Chongjian Ge, Shuchen Xue, Shaoteng Liu, Mengwei Ren, Soo Ye Kim, Yuqian Zhou, Qing Liu, Daniil Pakhomov, Kai Zhang, Zhe Lin, and Ping Luo. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing, 2025. URL https://arxiv.org/abs/2512.17909.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders, 2025. URL https://arxiv.org/abs/2510.11690.

Xingyu Zhou, Qifan Li, Xiaobin Hu, Hai Chen, and Shuhang Gu. Guiding a diffusion transformer with the internal dynamics of itself, 2026. URL https://arxiv.org/abs/2512.24176.

Chenchen Zhu, Saksham Suri, Cijo Jose, Maxime Oquab, Marc Szafraniec, Wei Wen, Yunyang Xiong, Patrick Labatut, Piotr Bojanowski, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficient universal perception encoder, 2026a. URL https://arxiv.org/abs/2603.22387.

Xuanyu Zhu, Yan Bai, Yang Shi, Yihang Lou, Yuanxing Zhang, Jing Jin, and Yuan Zhou. Beyond the last layer: Multi-layer representation fusion for visual tokenization, 2026b. URL https://arxiv.org/abs/2605.10780.

APPENDIX CONTENTS

A Extended Related Work **15**

A.1 Representation autoencoders and tokenizer design 15

A.2 Representation alignment and guidance 15

A.3 Layer-fusion learning and stochastic configuration 16

B Experimental Setup and Evaluation **16**

B.1 Encoder, layer fusions, and evaluation metrics 16

B.2 Training budgets and regularization grids 16

B.3 Baselines and controlled comparisons 17

C Supplementary Reconstruction Results **17**

C.1 Decoder-rate selection across layer fusions 17

C.2 Qualitative reconstruction under changed fusions 17

C.3 Layer reliance, subset value, and marginal gains 18

D Supplementary Generation Results **19**

D.1 Decoder replacement with fixed generated latents 19

D.2 Complementarity of decoder and generator regularization 20

D.3 Rate selection under internal guidance 20

E Generalizing to Other Pretrained Encoder Families **21**

E.1 Reconstruction robustness across layer subsets 21

E.2 Generation gains from joint regularization 21

F Theoretical Analysis and Proofs **22**

F.1 Mean preservation and disagreement covariance 22

F.2 Stage-specific objectives and optimal linear predictors 24

F.3 Coverage of nonempty subsets 25

F.4 Separation from deterministic global fusion 26

F.5 Per-layer rates continuously relax subset selection 27

F.6 From exact identities to nonlinear-model intuition 28

A Extended Related Work

We organize prior work by the component it changes. The first group develops representation autoencoders and reconstruction-oriented tokenizers; the second changes how representations are aligned, combined across depth, or sampled during training.

A.1 Representation autoencoders and tokenizer design

Image representation autoencoders. Latent diffusion (Rombach et al., 2022) and diffusion transformers (Peebles & Xie, 2023) standardized training a diffusion model inside the latent space of a separately trained autoencoder rather than on raw pixels. In the standard pipeline, the tokenizer is a reconstruction-centric VAE. RAE (Zheng et al., 2025) instead pairs a frozen self-supervised vision encoder with a learned decoder, producing a semantically structured latent that remains reconstructable. RAEv2 (Singh et al., 2026) fuses the last k encoder layers and shows that k controls a reconstruction–generation Pareto frontier. Our work takes this result as its starting point: a fixed-k fusion asks one latent to serve two consumers with different preferences.

Video extensions. Recent work carries representation autoencoding beyond images. VideoRAE extracts multi-scale hierarchical features from frozen video foundation models and compresses them with a lightweight self-attention projector, producing continuous latents for diffusion transformers and discrete tokens for autoregressive models (Xie et al., 2026). V-RAE constructs compact video latents from frozen vision-foundation features using temporal pooling, and evaluates them for reconstruction, semantic probing, generation, and future prediction (Guo et al., 2026). These studies extend the scope of RAEs to temporally structured data; their focus is video compression and temporal modeling, whereas FuseReg targets robustness to alternative layer fusions from a fixed image encoder.

Reconstruction-focused adaptation and efficiency. Several works reduce the reconstruction gap by modifying the encoder, latent geometry, or training recipe. DINO-Tok (Jia et al., 2026) adapts DINO for visual tokenization; the Hyperspherical Autoencoder (Chang et al., 2026), VFM-VAE (Bi et al., 2026), and AlignTok (Chen et al., 2026) reshape or align foundation-model features for reconstruction and generation; and RPiAE (Gong et al., 2026) pivots the representation for both generation and editing. The Prism Hypothesis (Fan et al., 2026) and PS-VAE (Zhang et al., 2025) directly study the tension between semantic content and pixel fidelity. Improving Reconstruction of RAE (Liu et al., 2026) changes the reconstruction recipe, whereas FlatDINO (Calvo-González & Fleuret, 2026) compresses the spatial representation into a much shorter token sequence for efficient diffusion. These directions are complementary to FuseReg, which freezes the encoder and trains one decoder across a broad distribution of sampled layer subsets.

A.2 Representation alignment and guidance

Representation alignment and internal guidance. REPA (Yu et al., 2025) distills a pretrained encoder's representation into an intermediate DiT layer via a projection head; REPA-E (Leng et al., 2025) back-propagates this alignment loss into the tokenizer itself, and Singh et al. (2025) isolate spatial structure, rather than global semantic content, as an important ingredient. RAEv2 finds that REPA complements RAE by improving intermediate spatial structure (Singh et al., 2026), whereas PixelREPA (Shin et al., 2026) shows that direct REPA can fail in pixel space. Together, these results indicate that the benefit of alignment depends on the architecture and the space in which alignment is applied. AutoGuidance uses a separately trained degraded model (Karras et al., 2024); internal guidance instead reuses intermediate features from the same diffusion model (Singh et al., 2026; Zhou et al., 2026). Both motivate making the decoder robust to the latent fusion presented downstream.

A.3 Layer-fusion learning and stochastic configuration

Learned multi-layer fusion. Deeply supervised networks attach prediction heads at multiple depths (Lee et al., 2015). More directly, two contemporaneous papers study which frozen encoder layers to use and how to combine them. DRoRAE (Zhu et al., 2026b) learns a depth-routed, energy-constrained fusion module on top of a frozen decoder, trained in three phases so the fusion adapts to the decoder's implicit distributional constraints; it reports gains in both rFID and gFID over a last-layer RAE baseline. The attentive multilayer probe (Ciernik et al., 2026) instead learns, per downstream task, an attention distribution over ViT layers for representation probing rather than generation. Both confirm our central premise that useful information is spread unevenly across layers and a fixed fusion underuses it. However, both converge on one learned combination rule per model or task. Our globally learned gate in Figure 1 collapses toward the shallowest, most pixel-aligned layer because that is what a pixel-reconstruction loss cheaply rewards. This observation motivates FuseReg, which trains over random layer subsets to support reconstruction across different fusions.

Stochastic configuration training. Dropout (Srivastava et al., 2014) originally regularizes by randomly zeroing units; nested (ordered) dropout (Rippel et al., 2014) instead retains a randomly sized prefix by dropping trailing units to recover an ordered, PCA-like basis in a linear autoencoder. Our analysis in §4 builds on this mechanism and contrasts its ordering with permutation symmetry. Elasticity along other network axes has been explored primarily for inference efficiency: stochastic depth (Huang et al., 2016) and LayerDrop (Fan et al., 2020) randomly drop residual blocks or Transformer layers at training time so that shallower sub-networks can be extracted at test time without retraining; slimmable networks (Yu et al., 2019; Yu & Huang, 2019) do the same along channel width; and Matryoshka Representation Learning (Kusupati et al., 2022) trains embeddings that stay accurate under nested dimensionality truncations. In every one of these, the extracted sub-configuration is smaller than the full model and the goal is a compute/quality trade-off at deployment time. FuseReg applies layer dropout to the choice of frozen encoder layers that form the latent for the decoder or generator, placing every nonempty layer subset in the training support. The probability-weighted guarantee in Corollary 1 and the empirical subset analysis in §C.3 characterize the fusion robustness.

B Experimental Setup and Evaluation

This section collects the protocols shared by the experiments in §5 and the supplementary results below. We evaluate fusion robustness first at the decoder, then at the full generation pipeline.

B.1 Encoder, layer fusions, and evaluation metrics

The primary experiments and the supplementary reconstruction and generation results in Sections C and D use a frozen, pretrained DINOv3-L encoder and ImageNet at 256×256 resolution. Experiments with other pretrained encoder families are presented separately in §E. The candidate pool contains $K=23$ transformer layers. The main reconstruction comparisons use the full $k=23$ fusion, the $k=7$ fusion over layers $\{\ell_{11}, \ell_{13}, \ldots, \ell_{23}\}$, and the single layer ℓ_{11}. FuseReg normalizes each sampled fusion by its number of retained layers (Equation (2)); fixed-fusion baselines retain their stated scale conventions.

We evaluate reconstruction with PSNR, SSIM, and rFID on 50k images, and generation with gFID and IS over 50k class-conditional samples using 50 Euler steps. The per-layer diagnostic in Figure 4 uses the same 10,000 held-out images for both decoders and retains RAEv2's fixed ℓ_{23} token-mean surrogate in every probe.

B.2 Training budgets and regularization grids

Decoder sweeps train a ViT decoder from scratch for 16 epochs with the pixel, perceptual, adversarial, and classifier-surrogate terms. Architecture, parameter count, optimization settings, and training budget are matched across reproduced decoder variants. The DiT grids use 40 training epochs with the same controls within each generator scale; the DiT-XL hidden size is 1440. Table B.1 summa-

rizes the rates. DiT-Base is evaluated without guidance; DiT-XL is evaluated both without guidance and with internal guidance at scale 1.78.

Table B.1: **Training budgets and FuseReg rate grids.** Decoder and generator rates are varied independently. In the generator studies, each trained DiT is evaluated with the corresponding decoder-rate grid.

Component	Epochs	Drop rates
Decoder	16	$p_{\mathrm{dec}} \in \{0, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95\}$
DiT-Base	40	$p_{\mathrm{dit}}, p_{\mathrm{dec}} \in \{0, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9\}$
DiT-XL	40	$p_{\mathrm{dit}} \in \{0, 0.5, 0.7, 0.9\}$; decoder rates as above

B.3 Baselines and controlled comparisons

Table 1 uses the official RAEv2 fixed-fusion checkpoints at $k{=}7$ and $k{=}23$. The rate sweeps instead use reproduced $p{=}0$ baselines matched to their respective protocols; comparisons are made within each table. In the decoder-swap study, the generator and sampled latents remain fixed while the decoder changes, isolating the effect of reconstruction training on generated images. In the joint grids, both training rates vary. Reported grid entries are point estimates; the four-setting comparison in §D.2 describes their complementarity at the selected configuration.

C Supplementary Reconstruction Results

All reconstruction results in this section use a frozen, pretrained DINOv3-L encoder.

C.1 Decoder-rate selection across layer fusions

Table C.1: **FuseReg controls decoder robustness across layer fusions.** Each row is one decoder trained for 16 epochs with rate p_{dec} and evaluated on 50k ImageNet-256 images using three input fusions. All runs retain the GAN and classifier-surrogate terms. The $p_{\mathrm{dec}}{=}0$ row is our reproduced full-fusion baseline. **Bold**/underline: best/second-best within each metric and fusion.

	Fusion $k{=}7$			Fusion $k{=}23$			Single layer ℓ_{11}		
p_{dec}	**PSNR ↑**	**SSIM ↑**	**rFID ↓**	**PSNR ↑**	**SSIM ↑**	**rFID ↓**	**PSNR ↑**	**SSIM ↑**	**rFID ↓**
0 (RAEv2)	12.54	0.372	16.108	27.10	0.808	**0.189**	15.81	0.503	3.392
0.05	19.78	0.518	2.826	**28.59**	**0.853**	<u>0.273</u>	20.76	0.557	1.881
0.1	20.52	0.550	1.718	<u>28.58</u>	<u>0.852</u>	0.285	21.33	0.586	1.209
0.3	22.05	0.612	0.762	28.48	0.849	0.294	22.90	0.657	0.576
0.5	22.93	0.646	0.655	28.43	0.848	0.294	23.98	0.695	0.522
0.7	23.47	0.665	0.634	28.20	0.842	0.322	24.58	0.714	0.493
0.9	<u>23.62</u>	<u>0.671</u>	<u>0.610</u>	27.60	0.827	0.415	<u>24.82</u>	<u>0.722</u>	<u>0.458</u>
0.95	**23.77**	**0.678**	**0.604**	27.52	0.826	0.421	**25.13**	**0.735**	**0.455**

Table C.1 separates full-fusion fidelity from robustness to reduced layer inputs. Full-fusion PSNR peaks at $p_{\mathrm{dec}}{=}0.05$, while the $k{=}7$ and ℓ_{11} inputs improve throughout the sweep. At $p_{\mathrm{dec}}{=}0.95$, one FuseReg decoder reaches 23.77, 27.52, and 25.13 dB across the three fusions. This is the checkpoint used in Table 1. The nearby $p_{\mathrm{dec}}{=}0.9$ setting retains slightly higher full-fusion PSNR, illustrating how the rate can be selected for the intended mix of inputs.

C.2 Qualitative reconstruction under changed fusions

Figure C.1 shows reconstructions of four random ImageNet-256 validation images under two fusions that differ from the full $k{=}23$ input. Under the $k{=}7$ fusion, the RAEv2$_{K=23}$ decoder collapses to low-contrast, nearly uniform outputs (10.8–14.4 dB), whereas FuseReg recovers object structure and color and slightly exceeds the RAEv2$_{K=7}$ decoder trained on this exact fusion (19.6–22.4 vs. 18.9–21.9 dB). Under the single layer ℓ_{11}, both fixed-fusion baselines are evaluated off their training

input and degrade to 14.5–17.2 dB, while the same FuseReg decoder remains at 20.2–23.3 dB. These examples follow the aggregate PSNR trend in Table 1: one FuseReg decoder handles both a sparse and a single-layer fusion without retraining.

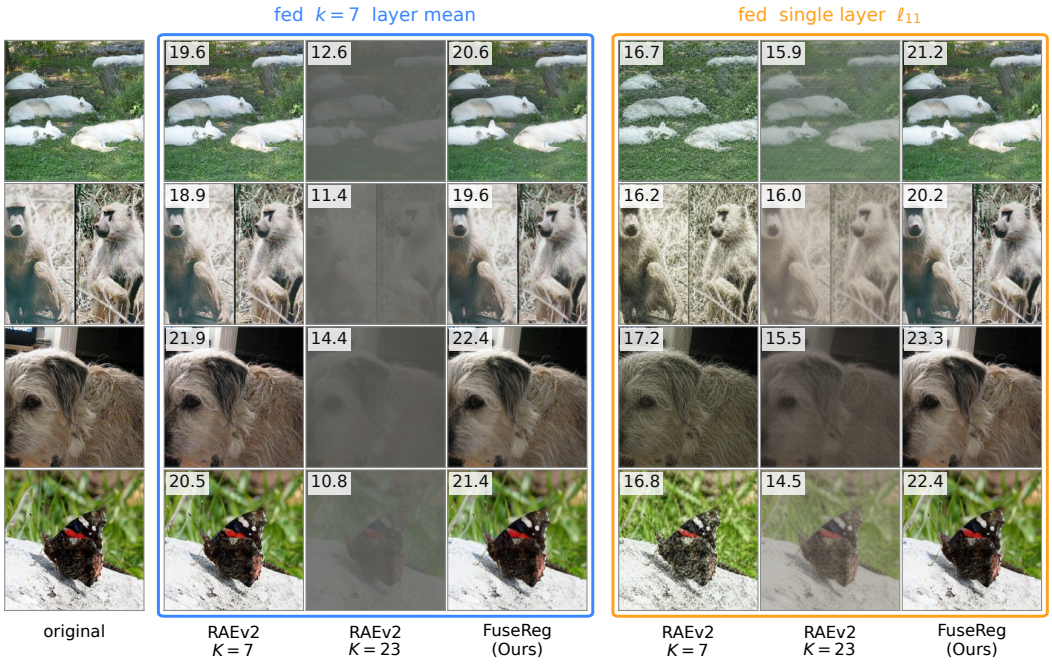


Figure C.1: **One FuseReg decoder supports both sparse and single-layer fusions.** Each row compares one image under the k=7 layers (blue) and ℓ_{11} (orange). RAEv2 labels denote the fixed training fusion: K=7 is matched in the blue block; both baselines are shifted in the orange block. The same FuseReg decoder preserves image structure in both settings. Insets show PSNR (dB).

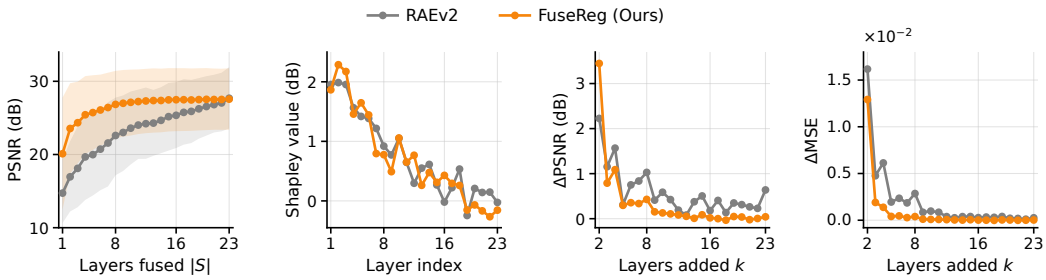
C.3 Layer reliance, subset value, and marginal gains

The probes in Figure 4 examine layer reliance and individual-layer readout. Removing one layer from the full fusion measures how strongly reconstruction depends on that layer in the presence of the others. Figure C.2 extends this analysis to combinations of layers. We sample random permutations of the 23 layers and evaluate their prefixes, using the same permutations and evaluation images for both decoders. Each prefix of length k defines a subset containing k layers.

Reconstruction quality from layer subsets. Figure C.2a measures how reconstruction quality changes with subset size. For each size, we aggregate per-image PSNR across the sampled subsets; lines show means and bands show the 10th–90th percentiles over image–subset pairs. The curve reveals how quickly each decoder approaches its full-fusion reconstruction quality as more layers become available.

Contributions of individual layers. Figure C.2b estimates each layer's Shapley value by averaging the PSNR change when that layer is added along sampled permutations. The empty-set value is defined by reconstruction from a zero latent input. This probe measures a layer's contribution across different subset contexts, complementing the removal probe, which evaluates its contribution specifically in the presence of all other layers.

Gains from increasing subset size. Figures C.2c and C.2d measure the marginal benefit of increasing subset size from $k-1$ to k. Let $\bar{v}(k)$ and $\bar{e}(k)$ denote mean PSNR and mean per-image MSE, respectively, aggregated over images and sampled subsets of size k. The two panels report $\bar{v}(k) - \bar{v}(k-1)$ and $\bar{e}(k-1) - \bar{e}(k)$. Here, k denotes subset size rather than an encoder layer index. Computing MSE reductions directly from reconstruction errors also checks whether diminishing gains persist beyond the logarithmic PSNR scale.



(a) Subset value (b) Shapley value (c) PSNR gain (d) MSE reduction

Figure C.2: **Subset diagnostics characterize reconstruction quality, layer contributions, and diminishing gains.** Frozen DINOv3 features, $K=23$; gray denotes RAEv2 and orange denotes FuseReg. **(a)** Reconstruction PSNR versus subset size; lines show means and bands the 10th–90th percentiles over evaluated image–subset pairs. **(b)** Monte-Carlo Shapley estimates attribute PSNR gains to individual layers across sampled permutations. **(c,d)** Differences between adjacent subset sizes, measured as increases in mean PSNR and reductions in mean MSE, respectively. FuseReg achieves strong reconstruction from small subsets and reaches diminishing returns earlier while broadly retaining the ordering of layer contributions.

FuseReg achieves higher reconstruction quality from small subsets and approaches its full-fusion performance with fewer layers. Its smaller marginal gains at larger subset sizes therefore reflect earlier saturation after a stronger initial readout. The corresponding decline in MSE reductions shows that this pattern also holds on the underlying error scale. Meanwhile, the broadly preserved Shapley ordering indicates that improved subset robustness coexists with unequal contributions across layers. Together, these diagnostics support the interpretation that FuseReg improves access to reconstruction information across depths. They characterize the nonlinear decoder empirically and complement the probability-weighted coverage analysis in Corollary 1.

D Supplementary Generation Results

All generation results in this section use a frozen, pretrained DINOv3-L encoder.

D.1 Decoder replacement with fixed generated latents

Table D.1 gives the full decoder-swap results underlying Figure 3. Each column pair keeps the DiT, fusion space, guidance setting, and sampled latents fixed. This isolates whether a FuseReg decoder improves the final images produced from an unchanged generator.

On the native $k=23$ fusion, increasing p_{dec} to 0.95 lowers unguided gFID from 3.01 to 2.21 and preserves guided gFID at 1.25. The $k=7$ columns test transfer away from the unregularized decoder's full-fusion training input: the same decoder replacement lowers gFID from 27.73 to 1.92 without guidance and from 16.35 to 1.42 with guidance. These comparisons show both an improvement within the original fusion space and reliable decoding after the generator's fusion space changes. Intermediate rates have different effects under guidance, motivating the joint-rate study below.

Table D.1: **FuseReg improves generation with a fixed generator.** Within each fusion and guidance setting, every decoder renders the same 50k latents from a fixed RAEv2 DiT (p_{dit}=0). The gray row uses the reproduced fixed-k=23 decoder; other rows vary p_{dec} over the same K=23 layer pool. Internal guidance uses scale 1.78 (Singh et al., 2026). **Bold**: best in each column.

	No guidance				Internal guidance (scale 1.78)			
	DiT k=23		DiT k=7		DiT k=23		DiT k=7	
Decoder p_{dec}	gFID↓	IS↑	gFID↓	IS↑	gFID↓	IS↑	gFID↓	IS↑
0.0 (RAEv2)	3.01	206.8	27.73	97.7	**1.25**	**263.8**	16.35	148.7
0.05	3.15	209.5	6.24	185.2	1.48	255.7	4.62	213.5
0.1	3.16	209.4	4.30	196.2	1.49	256.7	3.15	226.1
0.3	2.92	211.6	2.68	209.4	1.41	257.4	1.76	242.3
0.5	2.82	210.4	2.73	207.7	1.44	253.3	1.93	236.1
0.7	2.68	211.4	2.42	213.1	1.43	253.6	1.86	238.7
0.9	2.35	216.9	1.96	220.8	1.27	259.4	1.46	248.7
0.95	**2.21**	**218.5**	**1.92**	**221.8**	**1.25**	260.7	**1.42**	**250.5**

D.2 Complementarity of decoder and generator regularization

The unguided grids in Table 2 test whether applying FuseReg to both stages yields complementary gains. On DiT-Base, the baseline gFID is 13.96. Decoder-only regularization at p_{dec}=0.9 improves it by 1.00, and generator-only regularization at p_{dit}=0.7 improves it by 1.87. Their combination reaches 9.93, an improvement of 4.03, which is 1.16 beyond the additive extrapolation of the two individual gains. The analogous IS contrast is 4.9. This four-setting comparison describes complementarity at the selected configuration; it is not a statistical interaction test.

The DiT-XL grid separates this effect from a universal rate prescription. Its lowest unguided gFID occurs with strong decoder regularization and p_{dit}=0, whereas its highest IS occurs when both rates are large (Tables 2c and 2d). Decoder robustness remains useful across the two scales, while the preferred generator rate depends on model capacity and the metric being optimized.

D.3 Rate selection under internal guidance

Table D.2 completes the DiT-XL comparison with internal guidance at scale 1.78. The two panels report the same configurations using gFID and IS, making their different rate preferences explicit.

Table D.2: **FuseReg rate preferences under internal guidance on DiT-XL.** Class-conditional ImageNet-256, K=23, hidden size 1440, 40 training epochs, and 50k samples with guidance scale 1.78. Rows vary p_{dit} and columns vary p_{dec}. Colors encode metric values on a shared scale within each panel: blue denotes better values (lower gFID or higher IS), and orange denotes worse values. **Bold**: best in each panel.

(a) Generation FID ↓.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0.0	0.05	0.1	0.3	0.5	0.7	0.9	0.95
0	1.57	1.66	1.66	1.59	1.57	1.54	**1.50**	1.51
0.5	1.67	1.88	1.88	1.77	1.71	1.66	1.60	1.59
0.7	1.67	1.94	1.94	1.81	1.74	1.68	1.60	1.58
0.9	1.66	1.93	1.93	1.78	1.72	1.65	1.55	1.53

(b) Inception Score ↑.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0.0	0.05	0.1	0.3	0.5	0.7	0.9	0.95
0	**262.6**	255.8	256.7	257.3	258.2	258.2	258.5	259.4
0.5	258.8	249.8	250.7	252.3	253.0	253.2	254.5	255.8
0.7	257.1	247.7	248.2	249.4	251.1	251.2	251.8	253.5
0.9	261.9	252.6	253.3	255.1	256.6	257.0	258.1	260.1

The lowest guided gFID is 1.50 at $(p_{\text{dec}}, p_{\text{dit}}) = (0.9, 0)$, compared with 1.57 at $(0, 0)$. For every generator rate in the grid, a strongly regularized decoder improves gFID relative to the unregularized decoder in that row. The highest IS, however, remains at $(0, 0)$, in contrast to the unguided DiT-XL result in Table 2d. Thus the guidance setting changes which rate pair is preferred by each metric. The distinction motivates calibrating guidance and FuseReg together, as discussed in §6.

E GENERALIZING TO OTHER PRETRAINED ENCODER FAMILIES

The reconstruction robustness and generation gains of FuseReg extend beyond DINOv3-L to SigLIP2-L (Tschannen et al., 2025) and EUPE-B (Zhu et al., 2026a). SigLIP2-L is a vision-language encoder with $K{=}23$ candidate layers; EUPE-B is a 12-block ViT-B/16 with $K{=}11$ candidate layers (blocks 1–11). Both encoders remain frozen. We train decoders for 16 epochs and DiT-Base generators for 40 epochs on ImageNet-256, using the same recipes as in the DINOv3-L experiments and rates $p_{\mathrm{dec}}, p_{\mathrm{dit}} \in \{0, 0.3, 0.6, 0.9\}$.

E.1 RECONSTRUCTION ROBUSTNESS ACROSS LAYER SUBSETS

Table E.1 evaluates each decoder on full, sparse, and single-layer inputs. For SigLIP2-L, the sparse fusion uses layers $\{11, 13, \ldots, 23\}$ ($k{=}7$), and the single-layer input is ℓ_{11}. For the EUPE-B encoder, the corresponding inputs use $\{5, 7, 9, 11\}$ ($k{=}4$) and ℓ_5.

Table E.1: **Reconstruction robustness of FuseReg across encoder families.** Each row is one decoder trained for 16 epochs at rate p_{dec} and evaluated on 50k ImageNet-256 images with sparse, full, and single-layer inputs. The encoder headings separate the two families and their different layer subsets; every fusion includes the same last-layer classifier surrogate for that encoder. Gray rows mark the full-fusion baselines ($p_{\mathrm{dec}}{=}0$). **Bold**/underline: best/second-best.

SigLIP2-L ($K{=}23$)	Fusion $k{=}7$			Fusion $k{=}23$			Single layer ℓ_{11}		
p_{dec}	**PSNR ↑**	**SSIM ↑**	**rFID ↓**	**PSNR ↑**	**SSIM ↑**	**rFID ↓**	**PSNR ↑**	**SSIM ↑**	**rFID ↓**
0 (RAEv2)	13.05	0.288	107.689	**27.92**	**0.829**	**0.295**	11.27	0.193	198.638
0.3	21.73	0.594	**0.809**	27.87	0.827	0.325	22.38	0.623	**0.744**
0.6	22.78	0.635	0.947	27.80	0.825	0.452	23.58	0.667	0.860
0.9	**23.30**	**0.651**	0.999	27.03	0.803	0.522	**24.17**	**0.684**	0.815
EUPE-B ($K{=}11$)	Fusion $k{=}4$			Fusion $k{=}11$			Single layer ℓ_5		
p_{dec}	**PSNR ↑**	**SSIM ↑**	**rFID ↓**	**PSNR ↑**	**SSIM ↑**	**rFID ↓**	**PSNR ↑**	**SSIM ↑**	**rFID ↓**
0 (RAEv2)	15.69	0.437	11.053	**25.89**	**0.757**	**0.338**	19.28	0.497	3.109
0.3	22.92	0.629	0.657	25.71	0.750	0.385	23.37	0.653	0.683
0.6	**23.14**	**0.638**	**0.636**	25.45	0.740	0.388	23.74	0.665	0.561
0.9	22.97	0.633	0.644	24.74	0.713	0.492	**23.81**	**0.669**	**0.530**

For both encoders, FuseReg substantially improves reconstruction from sparse and single-layer inputs. A moderate decoder rate already yields large gains on these reduced inputs while preserving full-fusion reconstruction quality. The preferred rate varies with the encoder, input fusion, and metric, reflecting the trade-off between full-fusion fidelity and reduced-input robustness.

E.2 GENERATION GAINS FROM JOINT REGULARIZATION

Table E.2 evaluates FuseReg on both encoder families under the same unguided generation protocol. For each encoder and generator rate, all decoders render the same 50k latents sampled with 50 Euler steps. We compute gFID against the 50k ImageNet-256 validation images and IS.

Table E.2: **FuseReg at both stages across encoder families.** Unguided generation on class-conditional ImageNet-256 with frozen SigLIP2-L (K=23, top) and EUPE-B (K=11, bottom) encoders, 16-epoch decoders, and 40-epoch DiT-Base generators; 50k samples and 50 Euler steps. Rows vary p_{dit}; columns vary p_{dec}, with sampled latents fixed within each row. Each green box marks that encoder's fixed-fusion baseline. Shading is relative to this baseline (Blue = better, Orange = worse), with intensity scaled within each panel. **Bold**: best per panel.

(a) SigLIP2-L, gFID ↓.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0	0.3	0.6	0.9
0	13.77	11.80	10.91	9.73
0.3	11.93	10.33	9.61	8.69
0.6	11.26	9.77	9.08	**8.26**
0.9	11.97	10.00	9.30	8.49

(b) SigLIP2-L, IS ↑.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0	0.3	0.6	0.9
0	100.8	107.9	111.8	116.7
0.3	108.1	115.3	117.9	122.9
0.6	112.7	119.5	123.0	126.6
0.9	111.3	120.4	123.7	**127.9**

(c) EUPE-B, gFID ↓.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0	0.3	0.6	0.9
0	12.17	10.44	9.68	8.52
0.3	11.10	9.56	8.88	**8.00**
0.6	11.22	9.62	8.95	8.13
0.9	13.97	11.13	10.31	9.43

(d) EUPE-B, IS ↑.

$p_{\text{dit}} \backslash p_{\text{dec}}$	0	0.3	0.6	0.9
0	101.8	108.0	111.0	115.4
0.3	108.7	114.9	117.8	122.3
0.6	109.8	116.7	120.2	**123.7**
0.9	102.9	113.8	117.0	121.2

The complementary gains observed with DINOv3-L (Tables 2a and 2b) extend to both encoder families. Moving right along each row of Table E.2, stronger decoder regularization lowers gFID and raises IS. Relative to fixed fusion, applying FuseReg jointly reduces gFID from 13.77 to 8.26 for SigLIP2-L and from 12.17 to 8.00 for EUPE-B. In both grids, the best gFID and IS occur with a strongly regularized decoder, while the preferred generator rate depends on the encoder and metric.

F Theoretical Analysis and Proofs

This section analyzes the size-normalized, nonempty random-subset aggregation used by FuseReg in Equation (2). We first derive properties of the sampling rule itself, then specialize to linear squared-loss consumers, compare randomized and deterministic fusion, and finally state how far these results extend to the nonlinear models used in practice.

Scope at a glance. The conditioned subset probabilities, latent mean and covariance, support bound, and per-layer relaxation for fixed subset losses are exact without assuming a linear decoder or generator; the objective decompositions and closed-form optimizers require homogeneous linear predictors and squared loss. They provide mechanism-level intuition for the nonlinear ViT decoder and DiT rather than quantitative predictions for them.

F.1 Mean Preservation and Disagreement Covariance

Fix $K \geq 2$ encoder-layer features $h_1(x), \ldots, h_K(x) \in \mathbb{R}^d$ with finite second moments, and define their all-layer mean and deviations

$$\bar{h}(x) = \frac{1}{K} \sum_{k=1}^{K} h_k(x), \qquad e_k(x) = h_k(x) - \bar{h}(x), \qquad \sum_{k=1}^{K} e_k(x) = 0. \tag{10}$$

For $p \in [0, 1)$, let every layer be kept independently with probability $q = 1 - p$, condition on the kept set S being nonempty, and write

$$z_S(x) = \frac{1}{|S|} \sum_{k \in S} h_k(x), \qquad V(x) = \frac{1}{K} \sum_{k=1}^{K} e_k(x) e_k(x)^\top. \tag{11}$$

The conditional probability of a particular nonempty subset configuration is

$$\pi_p(S) = \frac{q^{|S|}p^{K-|S|}}{1-p^K}. \tag{12}$$

Let $R = |S|$ under this conditioned law and recall $c_K(p) = \mathbb{E}[(K-R)/(R(K-1))]$ from Equation (5).

Proof of Lemma 1. Conditioned on $R = r$, the kept set is uniform among all size-r subsets. Therefore

$$\Pr(k \in S \mid r) = \frac{r}{K}, \qquad \Pr(j,k \in S \mid r) = \frac{r(r-1)}{K(K-1)} \quad (j \neq k). \tag{13}$$

The conditional mean is

$$\mathbb{E}_S[z_S \mid r] = \frac{1}{r}\frac{r}{K}\sum_k h_k = \bar{h}. \tag{14}$$

For the conditional covariance, put $A_e = \sum_k e_k e_k^\top$ and use $\sum_{j\neq k} e_j e_k^\top = -A_e$:

$$\begin{aligned}\mathbb{E}_S[(z_S-\bar{h})(z_S-\bar{h})^\top \mid r] &= \frac{1}{r^2}\left(\frac{r}{K}A_e + \frac{r(r-1)}{K(K-1)}(-A_e)\right) \\ &= \frac{K-r}{r(K-1)}\,V(x).\end{aligned} \tag{15}$$

Let $\pi_r = \Pr(R=r)$ under the conditioned law above. By the law of total expectation,

$$\begin{aligned}\mathbb{E}_S[z_S] &= \sum_{r=1}^{K}\pi_r\,\mathbb{E}_S[z_S \mid r] \\ &= \sum_{r=1}^{K}\pi_r\,\bar{h} = \bar{h}.\end{aligned} \tag{16}$$

Similarly, for the centered second moment,

$$\begin{aligned}\mathbb{E}_S[(z_S-\bar{h})(z_S-\bar{h})^\top] &= \sum_{r=1}^{K}\pi_r\,\frac{K-r}{r(K-1)}V(x) \\ &= \underbrace{\mathbb{E}_R\left[\frac{K-R}{R(K-1)}\right]}_{c_K(p)}V(x).\end{aligned} \tag{17}$$

These identities give Equation (6). □

Remark 1 Drop rate controls disagreement variance

Writing $q = 1-p$ explicitly,

$$c_K(p) = \frac{\sum_{r=1}^{K}\binom{K}{r}q^r p^{K-r}\frac{K-r}{r(K-1)}}{1-p^K}. \tag{18}$$

Thus $c_K(0) = 0$, $c_K(p)$ is analytic for $0 < p < 1$, and $c_K(p) \to 1$ as $p \to 1$. Moreover, the conditioned subset size R decreases in the usual stochastic order as p increases, while $(K-R)/(R(K-1))$ decreases with R; hence $c_K(p)$ is nondecreasing. Thus the bounded coefficient $c_K(p) \in [0,1]$ increases covariance along layer-disagreement directions while leaving the expected latent equal to $\bar{h}$.

Insight. Increasing the drop rate makes the sampled representation more variable where layers disagree, while its mean stays fixed. This gives a single control for the strength of the perturbation, with a bounded effect on its covariance.

F.2 Stage-specific objectives and optimal linear predictors

Let $y(x)$ have finite second moment and define

$$\Sigma_{\bar{h}} = \mathbb{E}_x[\bar{h}\bar{h}^\top], \qquad \Gamma = \mathbb{E}_x[V(x)], \qquad A = \mathbb{E}_x[y\bar{h}^\top]. \tag{19}$$

Proof of the decoder identity. Write $z_S = \bar{h} + \delta_S$. By Lemma 1,

$$\mathbb{E}_S[\delta_S \mid x] = 0, \qquad \mathbb{E}_S[\delta_S\delta_S^\top \mid x] = c_K(p)V(x). \tag{20}$$

Conditioning on x and expanding the square gives

$$\begin{aligned}\mathbb{E}_S\|Dz_S - y\|^2 &= \mathbb{E}_S\|D\bar{h} - y + D\delta_S\|^2 \\ &= \|D\bar{h} - y\|^2 + 2(D\bar{h} - y)^\top D\,\mathbb{E}_S[\delta_S \mid x] \\ &\quad + \mathbb{E}_S[\delta_S^\top D^\top D\delta_S \mid x] \\ &= \|D\bar{h} - y\|^2 + c_K(p)\,\mathrm{tr}\big(DV(x)D^\top\big).\end{aligned} \tag{21}$$

The cross term vanishes by the first identity in Equation (20). For the quadratic term, use $u^\top Bu = \mathrm{tr}(Buu^\top)$ and then the second identity. Averaging Equation (21) over x replaces $V(x)$ by Γ and proves the decoder part of Proposition 1. Finally,

$$\mathrm{tr}(D\Gamma D^\top) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_x\,\mathrm{tr}(De_ke_k^\top D^\top) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_x\|D(h_k - \bar{h})\|^2, \tag{22}$$

which gives the stated form of $\Omega(D)$. □

Closed-form decoder. Expanding the resulting quadratic objective in D gives

$$\mathcal{L}_p^{\mathrm{dec}}(D) = \mathrm{tr}\big[D(\Sigma_{\bar{h}} + c_K(p)\Gamma)D^\top\big] - 2\,\mathrm{tr}(AD^\top) + \mathbb{E}\|y\|^2. \tag{23}$$

If $\Sigma_{\bar{h}} + c_K(p)\Gamma$ is positive definite, differentiating with respect to D gives the unique minimizer

$$D_p^\star = A\big(\Sigma_{\bar{h}} + c_K(p)\Gamma\big)^{-1}. \tag{24}$$

The disagreement functional is zero exactly when $D(h_k - \bar{h}) = 0$ almost surely for every k; the closed form suppresses directions with large disagreement covariance.

Time weighting and x-prediction. RAEv2 parameterizes the model output as a clean-latent prediction and computes its squared error after conversion to velocity (Singh et al., 2026). Let τ denote the implementation's data-to-noise time and set $t = 1 - \tau$ to match our noise-to-data convention. The training time law ν is defined by

$$u \sim \mathcal{N}(0,1), \qquad s = \mathrm{sigmoid}(u), \qquad \tau = \frac{as}{1 + (a-1)s}, \qquad t = 1 - \tau, \tag{25}$$

with shift $a = 8$ for our 256×1024 latents. Write $d_\tau = \max(\tau, 0.05)$ and $\rho(t) = d_\tau^{-2}$. For a subset input $z_{t,S}$ and full-layer target $\bar{h}$, the implemented loss is

$$\left\|\frac{z_{t,S} - x_\theta}{d_\tau} - \frac{z_{t,S} - \bar{h}}{d_\tau}\right\|^2 = \rho(t)\|x_\theta - \bar{h}\|^2, \tag{26}$$

where x_θ denotes the clean-latent prediction at this input and time.

Proof of the DiT identity. For fixed $t \in (0,1]$, write

$$z_{t,S} = t(\bar{h} + \delta_S) + (1-t)\epsilon = \bar{z}_t + t\delta_S, \qquad \bar{z}_t = t\bar{h} + (1-t)\epsilon. \tag{27}$$

Condition on (x, ϵ). Because the subset is sampled independently of ϵ, the conditional first and second moments of δ_S are still given by Equation (20). Therefore

$$\begin{aligned}\mathbb{E}_S\|Wz_{t,S} - \bar{h}\|^2 &= \|W\bar{z}_t - \bar{h}\|^2 + 2t(W\bar{z}_t - \bar{h})^\top W\mathbb{E}_S[\delta_S \mid x] \\ &\quad + t^2\mathbb{E}_S[\delta_S^\top W^\top W\delta_S \mid x] \\ &= \|W\bar{z}_t - \bar{h}\|^2 + t^2 c_K(p)\,\mathrm{tr}\big(WV(x)W^\top\big).\end{aligned} \tag{28}$$

Averaging over (x, ϵ) and multiplying by the deterministic weight $\rho(t)$ proves the DiT part of Proposition 1. The factor t^2 appears because subset variation is multiplied by t before entering the linear predictor; $\rho(t)$ preserves the velocity-space weighting of the x-prediction error. □

Proposition 3 Closed-form linear consensus predictor

For $t \in (0, 1]$, assume $\Sigma_{\bar{h}} + c_K(p)\Gamma + \sigma_t^2 I$ is invertible (which holds automatically for $0 < t < 1$). The optimal homogeneous linear predictor of $\bar{h}$ from $z_{t,S}$ under the weighted loss in Equation (8) is

$$W^\star_{p,t} = \frac{1}{t}\,\Sigma_{\bar{h}}\left(\Sigma_{\bar{h}} + c_K(p)\Gamma + \sigma_t^2 I\right)^{-1}, \qquad \sigma_t^2 = \frac{(1-t)^2}{t^2}. \tag{29}$$

If $\Sigma_{\bar{h}}$ and Γ share an eigenvector with eigenvalues (λ, γ), the corresponding gain, apart from the common factor $1/t$, is

$$\frac{\lambda}{\lambda + c_K(p)\gamma + \sigma_t^2}. \tag{30}$$

Along a shared eigendirection with nonzero signal ($\lambda > 0$), increasing the drop rate suppresses the predictor gain whenever the layers disagree ($\gamma > 0$), while a consensus direction ($\gamma = 0$) is unaffected by the subset-induced penalty.

Insight. The formula balances useful signal against noise and layer disagreement. Along the shared directions considered here, a higher drop rate reduces reliance on features that vary across layers. This explains how subset training favors information that survives changes in the chosen layers.

Proof. At fixed t, the positive scalar $\rho(t)$ cancels from the normal equation for predicting $\bar{h}$ from $z_{t,S}$:

$$W\,\mathbb{E}[z_{t,S} z_{t,S}^\top] = \mathbb{E}[\bar{h} z_{t,S}^\top]. \tag{31}$$

Mean preservation and Equation (20) imply

$$\mathbb{E}[\bar{h} z_S^\top] = \Sigma_{\bar{h}}, \qquad \mathbb{E}[z_S z_S^\top] = \Sigma_{\bar{h}} + c_K(p)\Gamma. \tag{32}$$

Because ϵ is independent, zero mean, and has identity covariance, all cross moments containing one factor of ϵ vanish. Hence

$$\mathbb{E}[\bar{h} z_{t,S}^\top] = t\Sigma_{\bar{h}}, \tag{33}$$

$$\mathbb{E}[z_{t,S} z_{t,S}^\top] = t^2\left(\Sigma_{\bar{h}} + c_K(p)\Gamma\right) + (1-t)^2 I. \tag{34}$$

For $0 < t < 1$, the positive noise term makes this second-moment matrix positive definite, since $\Sigma_{\bar{h}}$ and $c_K(p)\Gamma$ are positive semidefinite. Substituting these moments into Equation (31), factoring out t^2, and using invertibility yields Equation (29). On a shared eigenvector, the matrix factor acts as scalar multiplication by $\lambda/(\lambda + c_K(p)\gamma + \sigma_t^2)$, which proves Equation (30). □

Both DiT output heads are trained against the full-layer target. For linear heads $W_{\text{full},t}$ and $W_{\text{base},t}$, with base-head coefficient $\beta = 1$ in our experiments, averaging their losses over $t \sim \nu$ gives the added penalty

$$c_K(p_{\text{dit}})\mathbb{E}_{t\sim\nu}\left[\rho(t)t^2\left\{\Omega(W_{\text{full},t}) + \beta\Omega(W_{\text{base},t})\right\}\right], \tag{35}$$

where $\Omega(W) = \operatorname{tr}(W\Gamma W^\top)$. The time law and weight therefore determine how strongly each time contributes to training, while the fixed-time closed form remains unchanged. The same base disagreement matrix Γ enters both stages: the decoder pays an invariance penalty against a pixel target, and the DiT pays a time-weighted penalty while predicting the full-layer consensus from a noisy subset input.

F.3 Coverage of nonempty subsets

Corollary 1 Probability-weighted control of every subset

Let $\ell_S(f) \geq 0$ denote the expected loss of any predictor f conditional on the kept set being S. For every $0 < p < 1$ and every nonempty S_0,

$$\ell_{S_0}(f) \leq \frac{\mathcal{L}_p(f)}{\pi_p(S_0)}, \qquad \mathcal{L}_p(f) = \sum_{S \neq \emptyset} \pi_p(S)\ell_S(f). \tag{36}$$

The statement applies to any nonnegative reconstruction loss and to the weighted squared DiT training loss conditioned on S_0 (including its expectations over t and ϵ).

Insight. FuseReg includes every nonempty fusion in its training objective. A small average loss therefore limits each subset's loss, with tighter control for subsets sampled more often. This explains the coverage supplied by randomization and why that coverage is strongest for common subsets.

Proof. Every term in the sum defining $\mathcal{L}_p(f)$ is nonnegative, so $\mathcal{L}_p(f) \geq \pi_p(S_0)\ell_{S_0}(f)$. □

We evaluate reconstruction across layer subsets in Figure C.2.

Proposition 4 A fixed fusion need not transfer to an adjacent subset

For any loss threshold $M > 0$, there is a scalar $K = 2$ squared-error problem on which a decoder has zero loss on the all-layer fusion but loss greater than M on a Hamming-adjacent singleton fusion.

Insight. Two layers can cancel each other's noise when fused, so perfect reconstruction from the full fusion can hide arbitrarily large errors after one layer is removed. This motivates training over a distribution of layer fusions, as in FuseReg, rather than relying on performance at a single fixed fusion.

Proof. Fix $\eta > 0$. Let $y \sim \mathcal{N}(0, 1)$ and $g \sim \mathcal{N}(0, \sigma^2)$ be independent, and set $h_1 = g$ and $h_2 = \eta y - g$. The normalized all-layer fusion is $\bar{h} = \eta y/2$, so $D_{\text{full}} = 2/\eta$ has zero loss. On $S = \{1\}$, its loss is $4\sigma^2/\eta^2 + 1$, which exceeds M for a suitable σ. □

F.4 Separation from deterministic global fusion

For a general input-independent selection rule, let $w \in \mathbb{R}^K$ be its random layer-weight vector and $z_w = \sum_k w_k h_k$. Define $M_1 = \mathbb{E}[w]$ and $M_2 = \mathbb{E}[ww^\top]$.

Lemma 2 The linear objective depends only on two weight moments

For a homogeneous linear consumer and squared error,

$$\mathcal{L}(D; M_1, M_2) = \operatorname{tr}\big(D\Sigma(M_2)D^\top\big) - 2\operatorname{tr}\big(B(M_1)D^\top\big) + \mathbb{E}_x\|y\|^2, \tag{37}$$

$$\Sigma(M_2) := \sum_{j,k} [M_2]_{jk}\, \mathbb{E}_x[h_j h_k^\top], \qquad B(M_1) := \sum_k [M_1]_k\, \mathbb{E}_x[y h_k^\top]. \tag{38}$$

Thus the selection rule enters the linear objective only through (M_1, M_2).

Insight. For a linear predictor with squared error, fusion rules matter through the average layer weights and how those weights vary together. Matching these quantities gives the same training objective on the same features. This lets us compare fusion rules without analyzing every subset separately.

Proof. Because the selection rule is input independent, w and x are independent. Expanding the square gives

$$\begin{aligned}\mathcal{L}(D) &= \mathbb{E}\left[\left\| D\sum_k w_k h_k - y\right\|^2\right] \\ &= \sum_{j,k}\mathbb{E}[w_j w_k]\,\mathrm{tr}\big(D\mathbb{E}_x[h_j h_k^\top]D^\top\big) \\ &\quad - 2\sum_k \mathbb{E}[w_k]\,\mathrm{tr}\big(\mathbb{E}_x[y h_k^\top]D^\top\big) + \mathbb{E}_x\|y\|^2. \end{aligned} \tag{39}$$

Substituting $[M_2]_{jk} = \mathbb{E}[w_j w_k]$ and $[M_1]_k = \mathbb{E}[w_k]$, then collecting the two sums into $\Sigma(M_2)$ and $B(M_1)$, yields Equation (38). □

Proof of Proposition 2. For normalized random subsets, put $P = I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$. The weight vector $w_S = \mathbf{1}_S/|S|$ is exchangeable and satisfies $\mathbf{1}^\top w_S = 1$, so $\mathbb{E}[w_S] = K^{-1}\mathbf{1}$. Exchangeability makes all diagonal covariance entries equal and all off-diagonal entries equal, while $\mathbf{1}^\top w_S = 1$ implies that the covariance annihilates $\mathbf{1}$. It must therefore have the form $\mathrm{Cov}(w_S) = \alpha P$ for some scalar α. To identify α, note that

$$\mathrm{tr}(\mathrm{Cov}(w_S)) = \mathbb{E}\left[\frac{1}{R}\right] - \frac{1}{K} = \frac{K-1}{K}c_K(p),$$

which gives $\alpha = c_K(p)/K$. Hence

$$\mathbb{E}[w_S] = \frac{1}{K}\mathbf{1}, \qquad \mathrm{Cov}(w_S) = \frac{c_K(p)}{K}P, \qquad M_2 = \frac{1}{K^2}\mathbf{1}\mathbf{1}^\top + \frac{c_K(p)}{K}P. \tag{40}$$

The vector $\mathbf{1}$ is an eigenvector of M_2 with eigenvalue $1/K$, while every vector orthogonal to $\mathbf{1}$ has eigenvalue $c_K(p)/K > 0$. Thus M_2 has rank K. The deterministic claim follows from the rank of an outer product. □

For a given feature distribution, Lemma 2 shows that two selection rules induce distinct objectives as functions of D exactly when their induced $\Sigma(M_2)$ or $B(M_1)$ differs. Thus Proposition 2 establishes separation of the weight moments; feature moments determine whether this separation persists in the objective. It covers fixed subsets and globally learned layer weights. It does not cover an input-dependent router, whose weights are correlated with x and therefore fall outside the independent selection-rule model used by Lemma 2.

F.5 Per-layer rates continuously relax subset selection

Let $r = (r_1, \ldots, r_K) \in [0,1]^K \setminus \{\mathbf{0}\}$ contain per-layer keep probabilities and define

$$Z(r) = 1 - \prod_{k=1}^K (1 - r_k), \qquad \pi_r(S) = \frac{\prod_{k\in S} r_k \prod_{k\notin S}(1-r_k)}{Z(r)} \quad (S \neq \emptyset). \tag{41}$$

For fixed subset losses $\ell(S)$, the exact conditioned Bernoulli extension is

$$F(r) = \sum_{\emptyset\neq S\subseteq[K]} \pi_r(S)\ell(S) = \frac{\sum_{\emptyset\neq S\subseteq[K]} \ell(S)\prod_{k\in S} r_k \prod_{k\notin S}(1-r_k)}{1 - \prod_k (1-r_k)}. \tag{42}$$

The numerator is multi-affine; conditioning makes F a rational rather than a multilinear function.

Recovery of deterministic subsets. For every nonempty $S_0 \subseteq [K]$,

$$F(\mathbf{1}_{S_0}) = \ell(S_0). \tag{43}$$

At this vertex, the conditional distribution places unit mass on S_0; hence the nonzero vertices of the per-layer rate hypercube recover every nonempty deterministic subset exactly.

Interior smoothness. The function F in Equation (42) is real-analytic on $(0,1)^K$ because its numerator and denominator are polynomials and the denominator is positive there. On the exchangeable line $r_1 = \cdots = r_K = 1 - p$, the moments in Lemma 1 and the optimal linear predictors in Equations (24) and (29) are analytic for $0 < p < 1$ whenever their second-moment matrices remain invertible: $c_K(p)$ is a ratio of polynomials with a positive denominator, and matrix inversion is analytic on nonsingular matrices.

The full K-dimensional rate vector is the relaxation whose nonzero vertices represent all nonempty subsets. FuseReg uses the exchangeable diagonal $r_1 = \cdots = r_K = 1 - p$ at each stage. Consequently $(p_{\mathrm{dec}}, p_{\mathrm{dit}})$ is a structured two-dimensional slice of the $2K$-dimensional product relaxation, not an exact parameterization of every subset pair.

F.6 From exact identities to nonlinear-model intuition

The subset distribution, mean-preservation and covariance identities in Lemma 1, and the full-support bound in Corollary 1 do not assume a linear consumer. Likewise, the vertex and analyticity statements in §F.5 are exact for fixed subset losses. By contrast, Propositions 1 and 3 assume homogeneous linear predictors and squared loss; the eigendirection interpretation additionally assumes shared eigenvectors. For a nonlinear consumer, latent mean preservation does not imply prediction mean preservation or an exact disagreement penalty. We therefore use the linear results as mechanism-level intuition and evaluate the actual ViT decoder, mixed reconstruction loss, nonlinear DiT, and generation metrics empirically. In particular, the analysis neither predicts the sign of the two-rate contrast nor establishes super-additivity.