DECODING REGISTER TOKENS PRESERVES GLOBAL SCENE DETAILS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained vision transformers learn structured image representations spanning spatial patch tokens and non-spatial tokens. Recent analyses suggest that non-spatial register tokens encode global scene properties such as illumination, blur, and style. Yet patch-based image decoders discard these outputs, leaving the decoder to recover image appearance from patches alone. We investigate whether retaining register features improves decoding when the complete spatial patch representation is already available. We study this question using representation autoencoders (RAEs) with frozen register-equipped encoders. Our decoder processes four register tokens together with all patch tokens through joint self-attention. Appearance probes, paired reconstruction-error analysis, and fixed-decoder register substitutions examine which global image properties are accessible from registers and how their values affect reconstructed pixels. For generation, we train a multi-head attention pooling (MHAP) predictor to recover register features from sampled patches, keeping the pretrained diffusion model and its samples unchanged. On ImageNet-1K, our ViT-B decoders achieve mean reconstruction FIDs of 0.472 with DINOv2 and 0.442 with DINOv3, reductions of 9.5% and 5.2% relative to their respective patch-only baselines. The corresponding PSNRs are 19.73 and 20.54 dB. For generation, our DINOv2 models achieve a mean FID of 1.641, 9.8% below the patch-only baseline on identical sampled patches.

1 INTRODUCTION

Representation autoencoders (RAEs) connect pretrained visual representations with latent image generation through a two-stage pipeline (Zheng et al., 2026; Tong et al., 2026). A decoder learns to reconstruct images from a frozen encoder, and a diffusion transformer (DiT) learns to generate features in that space. At inference, sampled features are decoded into pixels. The decoder must therefore use a representation shaped by the encoder's pretraining objective, rather than learned jointly for reconstruction.

Patch-based RAEs retain spatial patch tokens while discarding non-spatial encoder outputs. For register-equipped DINOv2 and DINOv3 encoders, this excludes four register features already computed in the same forward pass (Oquab et al., 2023; Darcet et al., 2024; Siméoni et al., 2025). Registers support non-local computation and reduce artifacts in spatial features (Darcet et al., 2024). More recently, *Into the Rabbit Hull* identified register-specialized concepts associated with illumination, blur, and style (Fel et al., 2026). Such properties matter to reconstruction: recovering recognizable objects does not ensure that their lighting, color relationships, or surrounding textures match the source image.

Discarding registers leaves the decoder to recover these properties from patches alone. Self-attention allows aggregation across all patches, but recoverability does not guarantee that a decoder learns an equally effective readout under its reconstruction objective and training budget. Image-reconstruction losses supervise the output, not recovery of the encoder's register state. Registers may therefore make appearance information easier to use even when related information remains in patches. Our question is: *does supplying register features improve image decoding when the complete spatial patch representation is retained?*

We prepend four register outputs to the patch sequence and decode both through joint self-attention, keeping the encoder frozen and introducing no additional decoder loss. On ImageNet-1K, DINOv2

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.

and DINOv3 ViT-B decoders achieve 9.5% and 5.2% lower reconstruction FID than their respective patch-only references. Without additional training, register-input decoding also improves rFID, PSNR, and DISTS on NINCO and Places365 for both encoder families. To examine the image properties involved, we combine appearance probes with a separate paired ImageNet-100 reconstruction analysis. The latter finds lower error in both mean RGB color and spatially centered images: the benefit extends beyond correcting a constant color offset. Fixed-decoder substitutions and interpolations further show that register values influence color and blur-related appearance while receiver patches remain unchanged. Together, these experiments examine both reconstruction fidelity and the decoder's response to changes in register state.

For generation, pretrained patch-latent DiTs do not produce register features. We train a multi-head attention pooling (MHAP) predictor (Lee et al., 2019) independently of the image decoder to recover registers from sampled patches. Within a fixed decoder, predicted registers outperform mean and shuffled replacements on encoded images, although encoder-produced registers retain the lowest reconstruction errors. Decoding identical DINOv2 patch samples with recovered registers achieves mean generation FID 1.641, compared with 1.819 for the patch-only reference, a 9.8% reduction. This improves the decoding of an existing generator's samples without retraining the DiT or adding register tokens to its diffusion process.

2 Related Work

Representation autoencoders and generative tokenizers. Latent diffusion models and diffusion transformers synthesize images in learned feature spaces (Rombach et al., 2022; Peebles & Xie, 2023). RAEs replace reconstruction-trained encoders with frozen vision backbones and learn decoders over their features (Zheng et al., 2026). Scale-RAE extends this approach to text-to-image generation (Tong et al., 2026), while RAEv2 aggregates features from multiple encoder layers and combines RAE latents with intermediate diffusion-feature alignment (Singh et al., 2026). LV-RAE adds a learned local-variation encoder to improve reconstruction (Liu et al., 2026). We retain all spatial patches and expose four register outputs from the same encoder, rather than adding another image encoder or modifying the representation modeled by diffusion.

Decoding pretrained representations. Image synthesis provides a way to examine which image properties a representation preserves. Diffusion Autoencoders learn semantic codes together with a diffusion decoder (Preechakul et al., 2022). RCDM visualizes self-supervised representations through conditional generation (Bordes et al., 2022), and unCLIP generates images from CLIP image embeddings (Ramesh et al., 2022). Pretrained representations also guide diffusion training through feature alignment and architectural changes, as in REPA and DDT (Yu et al., 2025; Wang et al., 2025). Our register-recovery pathway instead operates after sampling, leaving pretrained diffusion weights and sampled patches unchanged.

Registers and global appearance. Darcet et al. (2024) introduced register tokens to reduce artifacts associated with non-local computation in vision-transformer patch features. Subsequent work examines whether registers require explicit training and how their role varies across architectures (Jiang et al., 2025; Baxevanakis et al., 2026). *Into the Rabbit Hull* analyzes concept organization in DINO representations and identifies register-specialized concepts associated with illumination, blur, and style (Fel et al., 2026). We study the consequences for reconstruction and appearance control by using register outputs as decoder inputs, without altering register computation inside the encoder.

3 Method

We extend patch-based RAEs (Zheng et al., 2026) to decode encoder registers alongside spatial patches, and recover registers from sampled patches for generation.

3.1 Encoder representations and decoder inputs

For an image x, a frozen register-equipped DINOv2 or DINOv3 encoder produces N patch features $Z_p(x) \in \mathbb{R}^{N \times d}$ and four register outputs $Z_r(x) \in \mathbb{R}^{4 \times d}$, where d is the feature dimension. Both

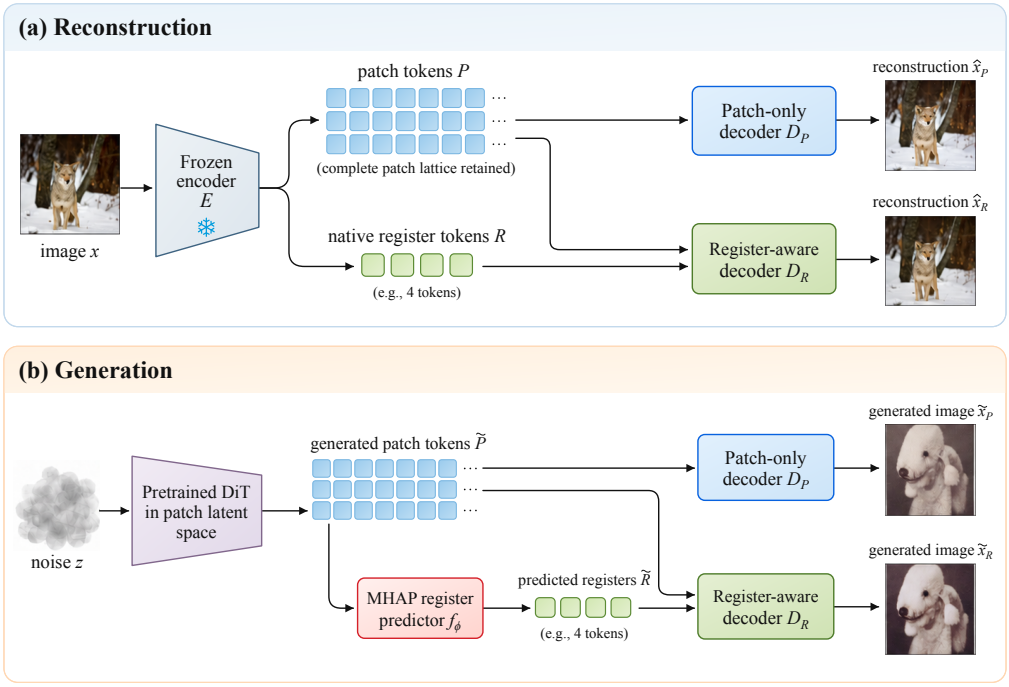

Figure 1: **Decoding patches and registers.** (a) The patch-only decoder D_P receives all spatial patches; D_R additionally receives four encoder registers. (b) MHAP recovers registers from sampled patches before decoding with D_R. Both generation routes use the same pretrained DiT and sampled patches.

decoding conditions use the same encoder and complete patch sequence; only the patch-and-register decoder receives the register outputs.

Let D_P and D_R denote the patch-only and register-input image decoders, with inputs expressed in encoder-feature coordinates. Their reconstructions are

$$\begin{aligned} \hat{x}_P &= D_P(Z_p(x)) \,, \\ \hat{x}_R &= D_R(Z_p(x), Z_r(x)) \,. \end{aligned} \tag{1}$$

Figure 1 summarizes reconstruction and generation.

3.2 Decoding patches and registers

Each decoder learns token-wise input projections. We denote the patch and register projections by W_p and W_r, respectively. The register decoder prepends the projected registers:

$$H^0 = \left[W_r Z_r + b_r;\, W_p Z_p + b_p\right]. \tag{2}$$

The semicolon denotes token concatenation. The decoder's own learned class token and positional embeddings are omitted from this expression; their conventions are specified in Appendix A.1.

Both decoder variants use a ViT-B transformer backbone and RGB reconstruction, perceptual, Gram, and adversarial image losses. The register-input decoder additionally receives four encoder registers through joint self-attention. For further details on decoder training we refer the reader to Appendix A.1. No additional register-prediction loss is used in decoder training. Section 4.3 fixes a trained decoder and its patches to test changes in register values.

3.3 Register recovery for generation

A pretrained patch-latent DiT does not produce registers. We recover them with a multi-head attention pooling (MHAP) predictor (Lee et al., 2019): four learned queries attend to all patches and predict

Table 1: **ImageNet-100 appearance probes.** Held-out R^2. DC measures spatial mean; AC contains nonzero low-frequency DCT coefficients.

Features	Thumbnail	Color hist.	DCT DC	DCT AC
Patch mean	0.211	0.288	0.735	0.038
Patch mean + registers	0.228	0.296	0.820	0.068

the four registers in encoder-defined order. Training uses image-matched patch–register pairs from the frozen encoder and mean squared error in standardized register coordinates, independently of the image decoder and diffusion model. The loss and coordinate conversions are specified in Appendix F.3.

Let $\widetilde{Z}_p$ denote sampled patches in encoder-feature coordinates and f_ϕ include conversion of predicted registers back to those coordinates. Generation uses

$$\widehat{Z}_r = f_\phi(\widetilde{Z}_p), \qquad \widetilde{x}_R = D_R(\widetilde{Z}_p, \widehat{Z}_r). \tag{3}$$

The patch-only reference decodes the same sampled patches with D_P. The register decoder and coordinate conversions remain fixed when native registers are replaced by predictions. Recovering registers once after sampling provides decoder features without additional diffusion targets; a separate joint-generation experiment appears in Appendix F.5.

4 Experiments

We test appearance information accessible from registers, their effect on reconstruction, and their recovery for generation. ImageNet-100 supports probes and paired appearance analysis; ImageNet-1K provides the principal reconstruction and generation evaluations.

Setup and evaluation. We use 12-block ViT-B decoders at 256×256 resolution; ImageNet-1K register decoders are trained for 16 epochs. Reconstruction and generation FID (rFID/gFID) compare Inception-feature distributions (Heusel et al., 2017). Preserving global scene details also requires fidelity to the source image's appearance, including color and brightness. For reconstruction, we therefore report PSNR for paired pixel agreement and DISTS for paired structure-and-texture discrepancy (Ding et al., 2021); higher PSNR and lower DISTS indicate closer reconstruction. Wavelet-distance results, metric protocols, and uncertainty definitions appear in Appendices G.5 and A.2.

4.1 Register state and global appearance

Ridge-regression probes predict an 8×8 RGB thumbnail, a color histogram, and low-frequency discrete cosine transform (DCT) coefficients from frozen DINOv2 features. The zero-frequency (DC) coefficient is proportional to each channel's spatial mean; the selected nonzero-frequency (AC) coefficients describe low-frequency spatial variation. We compare mean-pooled patches $p = \mathrm{mean}(Z_p)$ with the same features augmented by the four-register bank $r = \mathrm{vec}(Z_r)$. For each target y, we measure the change in held-out coefficient of determination, R^2:

$$\Delta R^2(r \mid p) = R^2([p; r] \to y) - R^2(p \to y). \tag{4}$$

Feature standardization and regularization selection use only the training portion of each split. Table 1 reports held-out prediction scores.

Adding registers improves all four prediction targets, with the largest increase for mean RGB color: DC R^2 rises from 0.735 to 0.820.

To test whether the benefit persists without averaging away patch positions, a separate probe retains the complete flattened patch sequence. Across three shared outer splits, adding native registers increases mean-color (DC) R^2 from 0.573 to 0.663 and thumbnail R^2 from 0.241 to 0.247. Aggregate non-DC prediction changes little ($\Delta R^2 = -1.9 \times 10^{-4}$). With identical splits and effective regularization



Figure 2: **Reconstruction from four registers alone.** Originals (top) and reconstructions from four registers without patches (bottom). Color and texture are partly recovered, while substantial object and spatial detail is missing. Decoder and selection details appear in Appendix B.3.

candidates, zero-register inputs reproduce the patch-only predictions exactly; shuffled and Gaussian ablation experiments provide no average gain on the DC or thumbnail targets (Appendix B.2).

These results show that registers expose global color in a form that the ridge readout can exploit alongside the patches. They characterize accessibility to this readout, rather than the total information available in the patch representation.

A decoder receiving only the four registers provides an image-space illustration: it recovers aspects of color and texture while losing substantial object and spatial detail (Figure 2). This behavior is consistent with the register specialization identified through dictionary learning in *Into the Rabbit Hull*, which associates register-specific concepts with illumination, motion blur, and style (Fel et al., 2026).

4.2 Where register inputs improve reconstruction

We compare patch-only and register-input DINOv2 reconstructions on the same 5,000 ImageNet-100 validation images for each of three decoder seeds. Register inputs increase PSNR by 0.473 dB and reduce DISTS by 0.007 relative to patch-only decoding, with paired-seed 95% intervals $[0.464, 0.481]$ dB and $[-0.008, -0.006]$, respectively. Both weighted DISTS components also decrease in every seed (Appendix G.1).

Mean color and spatial variation. To locate the pixel-error reduction, we decompose RGB MSE into a channel-mean term and a spatially varying term. Let $\mu(x)$ denote the vector of channel means and $x^c = x - \mu(x)$. Then

$$\mathrm{MSE}(x, \hat{x}) = \frac{1}{3} \|\mu(x) - \mu(\hat{x})\|_2^2 + \mathrm{MSE}(x^c, \hat{x}^c). \tag{5}$$

The first term measures errors in average RGB color. The second measures errors after centering the source and reconstruction separately by their own channel means. Centered values remain floating point and are not clipped.

Mean RGB MSE decreases by 8.52%. Of the absolute reduction, 43.21% occurs in the channel-mean term and 56.79% in the centered term. Centered MSE decreases by 5.28% on average, although it increases in 21.63% of image–seed pairs. The average benefit therefore extends beyond correcting image-wide color offsets to improving spatially varying reconstruction, with adverse cases retained in the analysis.

Residual-spectrum and wavelet-distance measurements appear in Appendix G. Figure 4(a) and Appendix H provide illustrative signed reconstruction-error maps for the ImageNet-1K models.

Decoder ablations. Registers yield lower rFID than CLS alone in both input studies; adding CLS to the register bank gives little further improvement (Appendix C.2). A separate CLS-plus-register comparison retains an rFID advantage over patch-only decoding across ViT-B, ViT-L, and ViT-XL (Appendix C.3). Learned-token and patch-summary ablations appear in Appendix C.5. On ImageNet-1K, native-register reconstruction and MHAP-based generation achieve lower FID than the two tested internal register-supervision variants (Appendix C.4).

4.3 Fixed-decoder register interventions

The reconstruction analysis identifies improvements in both mean color and spatial variation. We next examine how a trained decoder responds to the register values themselves: which aspects of reconstructed appearance can change when the complete patch representation remains fixed?

We first use a controlled appearance transformation. For each image, we extract features from the original and from a version modified by a combined color-and-saturation transformation T, obtaining (P, R) and (P_T, R_T). With decoder weights fixed, we evaluate all four combinations:

$$D_R(P, R), \quad D_R(P, R_T), \quad D_R(P_T, R), \quad D_R(P_T, R_T). \tag{6}$$

This compares native and substituted registers with either the original patches P or the transformed patches P_T held fixed. Register values are substituted in encoder-feature coordinates; normalization, token positions, and output processing remain unchanged.

With original patches fixed, substituting the transformed registers moves reconstructed mean RGB toward the transformed image. The mean directional coefficient is 0.499, with a paired-seed 95% interval of $[0.475, 0.524]$. This statistic measures the reconstruction's mean-color response along the applied transformation direction; it does not attribute a fraction of the patch-versus-register reconstruction gain to register values.

The response also extends beyond that direction. The orthogonal RGB-change norm is 0.0155, and classifier-logit cosine similarity to the original image decreases from 0.929 to 0.924. Thus, the decoder uses register values to alter reconstructed appearance, with accompanying off-target changes. The transformation and response statistics are defined in Appendix G.3.

Register interpolation. The controlled substitution transfers registers between two versions of the same image. We next explore what happens when the donor or register prompt is a different scene, tracing the reconstruction between the receiver's native register state and the donor's state.

For receiver registers R and donor registers $R^\star$,

$$R_\alpha = \mathrm{SLERP}(R, R^\star; \alpha), \qquad \hat{x}_\alpha = D_R(P, R_\alpha), \quad \alpha \in [0, 1]. \tag{7}$$

Corresponding slots share α; receiver patches, decoder weights, and positions remain fixed. Figure 3 shows two of 20 retained directed trajectories. Across all 20, endpoint changes follow the donor-defined direction in 19/20 cases for luminance and 20/20 for normalized gradient energy; 14/20 and 17/20 trajectories, respectively, are monotonic over the nine sampled values. Direction is defined by the donor's versus receiver's native reconstruction (Appendix G.4).

4.4 ImageNet-1K reconstruction

We next evaluate reconstruction at ImageNet-1K scale with both DINOv2 and DINOv3. Each decoder reconstructs the same 50,000 validation images (Table 2). Register-input models achieve 9.47% lower mean rFID with DINOv2 and 5.24% lower with DINOv3. For each encoder, all three register-input runs score below all three patch-only runs.

The distributional improvements are accompanied by better pixel agreement with the source images. Register-input decoders achieve mean PSNRs of 19.73 and 20.54 dB for DINOv2 and DINOv3, respectively, and reduce mean RGB reconstruction MSE by 3.28% and 2.96% relative to their patch-only references. MSE is lower for every decoder seed for both encoder families. The corresponding DISTS scores are 0.1044 and 0.0967; complete paired-fidelity measurements appear in Table 13 in the appendix.

Figure 4(a) illustrates the DINOv2 reconstruction differences, with signed error maps showing where register inputs improve or worsen pixel agreement. Additional examples appear in Figure 6.

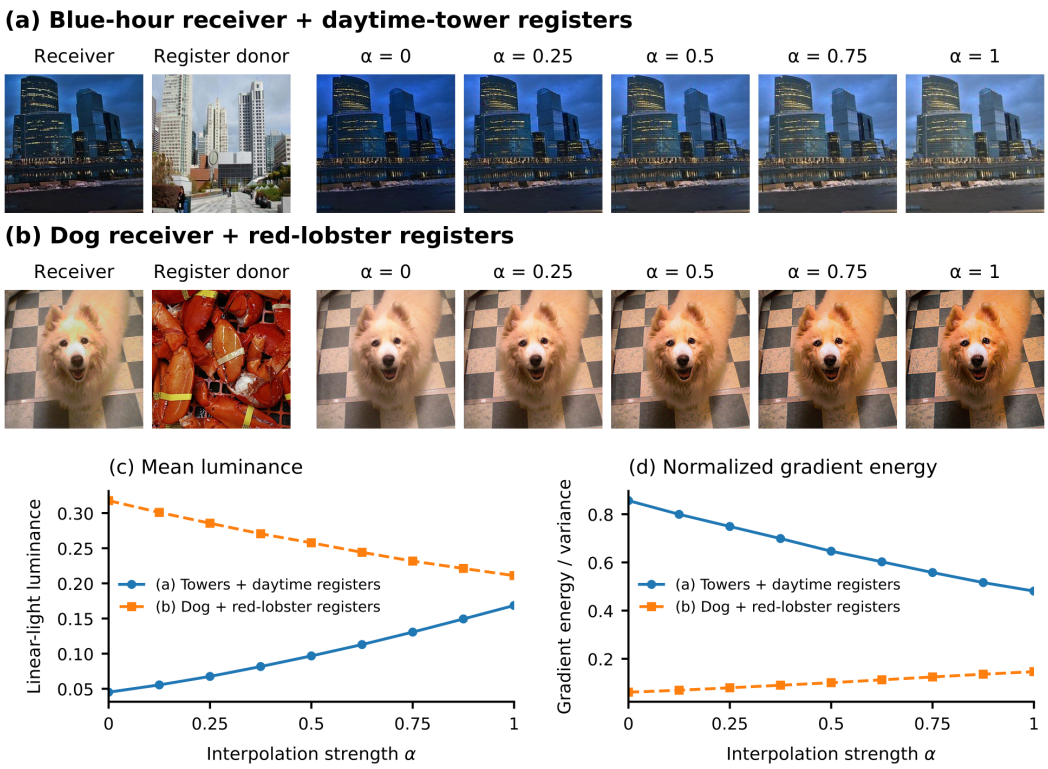

Figure 3: **Appearance changes through register interpolation.** (a) A blue-hour city receiver with registers from daytime towers. (b) A dog receiver with registers from a red-lobster image. Within each row, the receiver supplies the same complete patch sequence throughout. Each of the four register slots is interpolated separately toward its corresponding donor slot, using a shared interpolation coefficient α. Decoder weights and token positions remain fixed. At $\alpha = 0$, decoding uses the receiver's native registers; at $\alpha = 1$, it uses the donor's registers. Five of nine evaluated interpolation values are displayed. (c,d) Mean linear-light luminance and variance-normalized gradient energy for the two individual trajectories. Both rows use an ImageNet-1K-trained DINOv2 ViT-B decoder.

Table 2: **ImageNet-1K reconstruction.** Mean ± sample standard deviation across three ViT-B decoder seeds. rFID measures distributional similarity; PSNR measures pixel agreement with the source image.

Encoder	Decoder input	rFID ↓	PSNR (dB) ↑
DINOv2	Patches	0.5211 ± 0.0043	19.4901 ± 0.0072
	Patches + registers	$\mathbf{0.4717} \pm 0.0032$	$\mathbf{19.7262} \pm 0.0012$
DINOv3	Patches	0.4669 ± 0.0026	20.2994 ± 0.0182
	Patches + registers	$\mathbf{0.4424} \pm 0.0014$	$\mathbf{20.5406} \pm 0.0156$

The same ImageNet-1K-trained decoders also transfer without retraining to NINCO and Places365 (Appendix E). In the seed-0 evaluation, register inputs reduce rFID by 6.18%/43.93% for DINOv2 and 5.18%/28.14% for DINOv3 on NINCO/Places365, respectively. These distributional gains coincide with improved paired fidelity: PSNR increases by 0.17–0.40 dB and DISTS decreases by 0.0049–0.0065 across the four comparisons.

4.5 Register Recovery and Image Generation

Native versus predicted registers. Before evaluating generation, we test register recovery on 50,000 encoded ImageNet-1K validation images. We hold one DINOv2 register-input decoder and

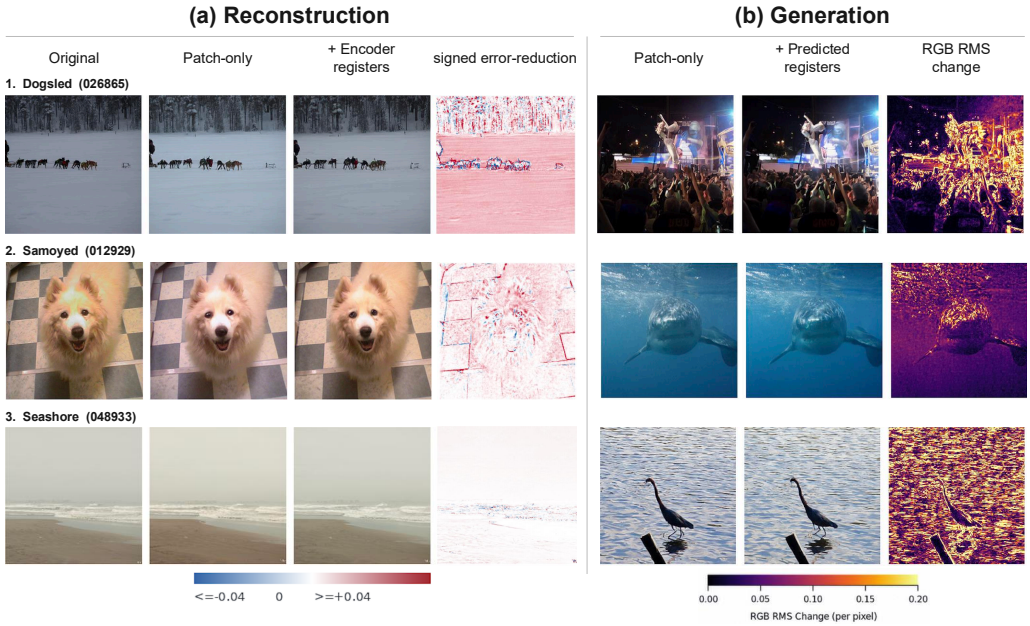

Figure 4: **Reconstruction and generation with register inputs.** (a) Original images, patch-only reconstructions, reconstructions with native encoder registers, and signed reconstruction-error reduction. Red indicates lower squared RGB error with registers; blue indicates lower error with patches alone. (b) Patch-only generation, generation with MHAP-predicted registers, and the unsigned per-pixel RGB RMS difference between these outputs. Each generation pair uses exactly the same stored patch tensor from the pretrained DiT-DH/XL. Generation difference maps show the magnitude of change, not improvement, because no paired ground-truth image exists. Reconstruction error maps use a fixed range of $[-0.04, 0.04]$; generation difference maps use $[0, 0.20]$. Examples are illustrative comparisons between separately trained decoders. Model and sample identities appear in Appendix F.6.

Table 3: **Decoding identical pretrained-XL samples.** ViT-B results are mean $\pm$ sample standard deviation across three decoder seeds, with one fixed MHAP predictor for register recovery. The released ViT-XL decoder is a larger-model reference.

Decoder	Decoding inputs	gFID ↓
ViT-B	Patches	1.8191 ± 0.0058
ViT-B	Patches + recovered registers	$\mathbf{1.6414} \pm 0.0133$
Released ViT-XL	Patches	1.5572

each image's complete patch sequence fixed, changing only the register source. DISTS is 0.1044 with native registers, 0.1121 with MHAP predictions, 0.1289 with the training-set mean, and 0.1333 with shuffled predictions. MHAP therefore recovers image-specific features useful for decoding, but native registers provide better reconstruction fidelity. Most of the RGB MSE increase when predictions replace native registers occurs in the channel-mean term (Appendix F.1).

Generation from identical sampled patches. We decode the same 50,000 DINOv2 patch samples from a released DiT-DH/XL using patch-only ViT-B, ViT-B with recovered registers, and the released ViT-XL decoder (Table 3). The generator, sampled tensors, and class labels are unchanged; Figure 4(b) shows paired decodings.

On identical pretrained-DiT patch samples, decoding with MHAP-predicted registers achieves mean gFID 1.6414, compared with 1.8191 for the patch-only reference, a reduction of 9.77%. All three register-decoder runs score below all three patch-only runs. The larger released decoder achieves

gFID 1.5572. This comparison uses unchanged diffusion samples, without another denoising pass. Generated images have no paired source, so PSNR and DISTS are not used as generation-quality scores.

5 DISCUSSION

Our analysis shows that register tokens encode global scene information. We showed this quantitatively in Section 4.1, where registers improved the prediction of global image properties such as color histograms in comparison to using mean-pooled patch tokens alone. We also showed it qualitatively by training a decoder only on the register tokens. The reconstructions from registers alone recover aspects of global appearance, such as the color and texture of grass or clouds, while losing substantial object and spatial detail (see Figure 2).

Current state-of-the-art decoders, such as those used in Representation Autoencoders (RAEs), take only the patch tokens of a pretrained encoder as input. We therefore tested whether the appearance information identified by our probing experiments also improves reconstruction quality when register tokens are given to the decoder as an additional input. Retaining four encoder registers alongside all spatial patches reduces mean ImageNet-1K reconstruction FID (rFID) by 9.5% with DINOv2 and 5.2% with DINOv3 relative to patch-only decoding, while also improving PSNR. The same decoders also improve out-of-distribution reconstruction without additional training: rFID decreases by 6.2% and 5.2% on NINCO and by 43.9% and 28.1% on Places365 for DINOv2 and DINOv3, respectively, alongside improved PSNR and DISTS.

During generation with RAEs, we recover the register tokens from the generated patch tokens using a trained MHAP predictor. This decouples the choice of decoder inputs from the choice of diffusion variables. For DINOv2, recovering registers from sampled patches reduces gFID by 9.8% relative to patch-only decoding, without changing the pretrained diffusion model or its samples. Registers can therefore improve image decoding without becoming additional diffusion targets. Our ablations show that correspondence matters: within-class shuffling preserves the register-bank distribution in each conditioning class but worsens gFID with the decoder fixed (Appendix F.2). The internal-supervision comparisons also favor the MHAP route over the tested alternatives, including the larger InfoNCE decoder (Appendix C.4).

While prior work has indicated that register tokens encode global image properties (Darcet et al., 2024; Fel et al., 2026), our global-appearance probes and qualitative reconstruction analysis offer further insights into the global properties that they encode, and why they may fall short of the CLS and patch tokens when probed for semantic class information (Baxevanakis et al., 2026). The improved classification performance reported by Darcet et al. (2024) may therefore stem from this separation: with global background information absorbed by the register tokens that are not used in the probe, the CLS and patch tokens retain more disentangled semantic class information and are less affected by background cues, which can act as confounders during classification (Lapuschkin et al., 2019). Registers may thus be dispensable for classification, but not in general. We show this for image decoding and generation in RAEs, where retaining registers improves both. Beyond this, registers can also be used directly for applications such as global scene style transfer through spherical interpolation between register tokens (Figure 3).

Future work could explore additional applications in which access to global appearance and scene information is valuable, such as domain generalization (Zhou et al., 2022), where background or image style often acts as a confounder. Another direction is exploring alternative architectures and loss objectives that integrate register tokens more directly into the decoder. Finally, analyzing which information each register stores, and whether more registers lead to better disentanglement or greater information content, would further deepen our understanding of their role.

AI USE STATEMENT

We used ChatGPT and Claude for assistance with conceptual framing, experimental planning, literature discovery, drafting, and LaTeX preparation. Codex assisted with model and evaluation implementation, debugging, configuration checks, and figure-export scripts. The authors take responsibility for the final manuscript, code, figures, and scientific claims.

REFERENCES

Spiros Baxevanakis, Platon Karageorgis, Ioannis Dravilas, and Konrad Szewczyk. Do all vision transformers need registers? a cross-architectural reassessment. *arXiv preprint arXiv:2603.25803*, 2026. URL https://arxiv.org/abs/2603.25803.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing ImageNet out-of-distribution detection evaluation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2471–2506. PMLR, 2023. URL https://proceedings.mlr.press/v202/bitterwolf23a.html.

Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*, 2022. URL https://arxiv.org/abs/2112.09164.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16588.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2020.3045810. URL https://arxiv.org/abs/2004.07728.

Thomas Fel, Binxu Wang, Michael A. Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep S. Lubana, Talia Konkle, Demba Ba, and Martin Wattenberg. Into the rabbit hull: From task-relevant concepts in DINO to minkowski geometry. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fa68d2dfd8eb6202e7902586197f7629-Abstract-Conference.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Nick Jiang, Amil Dravid, Alexei A. Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.08010.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. URL https://cs.stanford.edu/people/jcjohns/eccv16/.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4. URL http://www.nature.com/articles/s41467-019-08987-4.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

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. *arXiv preprint arXiv:2602.08620*, 2026. URL https://arxiv.org/abs/2602.08620.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

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. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Preechakul_Diffusion_Autoencoders_Toward_a_Meaningful_and_Decodable_Representation_CVPR_2022_paper.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. URL https://arxiv.org/abs/2204.06125.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

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. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

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

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

Lokesh Veeramacheneni, Moritz Wolter, Hilde Kuehne, and Juergen Gall. Fréchet wavelet distance: A domain-agnostic metric for image generation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6605c7d1c1ae1521b328d036df5be1ae-Abstract-Conference.html.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025. URL https://arxiv.org/abs/2504.05741.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d9e42b4d7163931f3689d6d6fbaa11d0-Abstract-Conference.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://richzhang.github.io/PerceptualSimilarity/.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c4141c12660ad3625eb4ae845e0a6f9-Abstract-Conference.html.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

A Experimental Details

A.1 Feature coordinates and reconstruction objective

The specifications below describe the, 16-epoch ImageNet-1K native-register decoders.

Frozen features and image preprocessing. The frozen checkpoint identifiers and revisions are:

- DINOv2: `facebook/dinov2-with-registers-base`, revision `a1d738ccfa7ae170945f210395d99dde8adb1805`.
- DINOv3: `facebook/dinov3-vitb16-pretrain-lvd1689m`, revision `5931719e67bbdb9737e363e781fb0c67687896bc`.

Their actual encoder inputs are 224×224 and 256×256, respectively, producing 16×16 patch grids with 768 channels and four ordered register tokens. We extract the final `last_hidden_state`, after the encoder's terminal LayerNorm. This LayerNorm retains channel normalization but has its learned scale and bias removed. Its variance is computed over the 768 channels of each token, with epsilon 10^{-6} for DINOv2 and 10^{-5} for DINOv3. The four registers retain their encoder order; the encoder CLS token is excluded from the native-register decoder inputs.

For decoder training, RGB images are resized with PIL bicubic interpolation to a shorter side of 384 pixels, rounding the other side down, and randomly cropped to 256×256. The crop is reproducible from the decoder seed, epoch, and image identity. Evaluation and normalization-statistic fitting use a fixed center crop: repeated BOX downsampling halves both dimensions while the shorter side is at least 512, followed by bicubic resizing to a shorter side of 256 (rounding each dimension to the nearest integer) and a centered 256×256 crop. Pixels are divided by 255. Before encoding, DINOv2 inputs are resized from 256 to 224 using tensor bicubic interpolation with `align_corners=False` and no antialiasing; DINOv3 retains the 256×256 input. Both encoders use channel means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$.

Coordinates and decoder positions. Let $P_{p,c}$ and $R_{j,c}$ denote terminal patch and register features. The receiving RAE's stored patch-coordinate transform is

$$S_P(P)_{p,c} = \frac{P_{p,c} - \mu^P_{p,c}}{\sqrt{v^P_{p,c} + 10^{-5}}}, \qquad S_R(R)_{j,c} = \frac{R_{j,c} - \mu^R_{j,c}}{\sqrt{v^R_{j,c} + 10^{-5}}}. \tag{8}$$

The means and population variances are fitted over all 1,279,690 training images, independently for every patch position and channel or register slot and channel. Thus patch statistics have shape $768 \times 16 \times 16$ and register statistics have shape 4×768. For each coordinate a, $\mu_a = N^{-1} \sum_n f_{n,a}$ and $v_a = N^{-1} \sum_n f^2_{n,a} - \mu^2_a$; accumulation uses FP64, stored statistics use FP32, and negative roundoff in v_a is clipped to zero. These statistics remain fixed during evaluation.

Patch latents are stored or passed through S_P and inverted before the decoder's patch projection. Native reconstruction passes the raw registers directly to the register projection, without S_R. MHAP target standardization uses its separately fitted predictor statistics, described with Equation 15. Its output is first converted back to encoder-feature coordinates. The generation wrapper applies S_R for the receiving RAE's interface, and the decode call immediately applies S_R^{-1}; the transformer therefore receives predicted registers in the same raw coordinates as native registers, up to floating-point roundoff. Generated patch tensors likewise undergo the source generator's inverse coordinate transform before the receiver's patch-coordinate conversion.

The decoder sequence contains its own learned class token, followed by four register tokens and then 256 patch tokens in raster order. This decoder class token is distinct from the omitted encoder CLS. Separate learned linear maps project patch and register features to width 768. Patch positions use fixed two-dimensional sinusoidal embeddings. If e_j denotes the stored learned positional vector for register slot j, its effective input is $W_R R_j + b_R + 2e_j$; e_j is initialized from $\mathcal{N}(0, 0.02^2)$ and is added in both auxiliary-token preparation and sequence-position addition. The same convention is used during training and inference. Reconstruction pixels are predicted only from the 256 patch positions.

Reconstruction objective. For a batch of RGB targets x and reconstructions $\hat{x}$, the image objective is

$$\mathcal{L}_{\text{image}} = \mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{perc}} + 0.05\,\mathcal{L}_{\text{Gram}} + w_{\text{adv}}(t)\mathcal{L}_{\text{adv}}. \tag{9}$$

Here $\mathcal{L}_{\text{RGB}} = \text{mean}\,|\hat{x} - x|$ averages over the batch, channels, and pixels. The perceptual term is frozen VGG16 LPIPS (Johnson et al., 2016; Zhang et al., 2018), using `relu1_2`, `relu2_2`, `relu3_3`, `relu4_3`, and `relu5_3` (feature indices $3, 8, 15, 22, 29$). For this term, the VGG input transform is $((2x-1)-s)/q$, where $s = (-0.030, -0.088, -0.188)$ and $q = (0.458, 0.448, 0.450)$. At every spatial position, each feature vector f is normalized as $f/(\sqrt{\sum_c f_c^2} + 10^{-10})$. The loss sums the five layers' squared feature differences weighted by the frozen pretrained LPIPS 1×1 channel projections, averages spatially within each layer, and averages over the batch. LPIPS is in evaluation mode, so its dropout is inactive.

For the Gram term, let $F_\ell(x) \in \mathbb{R}^{C_\ell \times H_\ell W_\ell}$ be flattened, unnormalized VGG16 activations at layers $\ell \in \{3, 8, 15\}$. Its VGG inputs are first clipped to $[0, 1]$ and normalized with the ImageNet channel means and standard deviations above. Define

$$G_\ell(x) = \frac{F_\ell(x)F_\ell(x)^\top}{C_\ell H_\ell W_\ell}, \qquad \mathcal{L}_{\text{Gram}} = \sum_{\ell \in \{3,8,15\}} \text{mean}\,|G_\ell(\hat{x}) - G_\ell(x)|\,. \tag{10}$$

Each mean is over batch members and all entries of the Gram matrix; the three layer terms are summed with unit weights. The RGB and LPIPS terms use the unclipped decoder output.

The decoder adversarial term is the negative mean discriminator logit on augmented $2\hat{x} - 1$. It activates after eight completed epochs. Writing $\mathcal{L}_0 = \mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{perc}} + 0.05\mathcal{L}_{\text{Gram}}$ and W_{out} for the final pixel-prediction matrix, its weight is

$$w_{\text{adv}}(t) = 0.75\,\mathbf{1}[t \geq 8]\ \text{stopgrad}\left[\text{clip}_{[0,1000]}\ \frac{\|\nabla_{W_{\text{out}}}\mathcal{L}_0\|_2}{\|\nabla_{W_{\text{out}}}\mathcal{L}_{\text{adv}}\|_2 + 10^{-6}}\right]. \tag{11}$$

There is no register-prediction term in the native-register decoder objective. MHAP is trained separately.

Optimizer and schedule. These decoders use AdamW with betas $(0.9, 0.95)$, weight decay zero, global batch size 256, BF16 computation, and no gradient clipping. Let u be the number of completed optimizer steps and $t = u/4998$ the epoch-equivalent training time. The learning rate, applied before each step, is

$$\eta(t) = \begin{cases} 2 \times 10^{-4}\,t, & 0 \leq t < 1, \\ 2 \times 10^{-5} + \frac{2\times10^{-4} - 2\times10^{-5}}{2}\left[1 + \cos\left(\pi\,\frac{t-1}{15}\right)\right], & 1 \leq t \leq 16. \end{cases} \tag{12}$$

Training uses 4,998 complete batches per epoch and 79,968 steps in total. EMA decay is 0.9978, and evaluation uses the final 16-epoch EMA export. Clean runs use zero patch noise, zero register noise, and zero register dropout. Evaluation outputs are clipped to $[0, 1]$, multiplied by 255, rounded, and stored as RGB uint8 images.

A.2 Evaluation and uncertainty

The ImageNet-100 appearance experiment evaluates 5,000 validation images for each decoder seed. ImageNet-1K reconstruction evaluates 50,000 validation images. Paired comparisons use the same image identities, preprocessing, and output conversion for both models.

For an image-level metric, we first average over the evaluation images within each decoder seed. We then compute the mean and sample standard deviation across decoder seeds. Paired confidence intervals are computed from the within-seed differences between models. With three seeds, the 95% interval is

$$\overline{\Delta} \ \pm\ t_{0.975,2}\frac{s_\Delta}{\sqrt{3}}, \tag{13}$$

where s_Δ is the sample standard deviation of the three seed-level differences. These intervals describe training-seed variation; evaluation images are not treated as independent training replicates. Tables retaining a separately reported standard-deviation convention identify it in their captions.

Table 4: **Decoder-stage inference cost.** Median CUDA-event latency per batch from 100 forwards after 20 warmups. Each route and batch size uses a fresh process; the two routes use identical stored GPU-resident patch inputs at each batch size. Peak allocated memory includes route weights, inputs, temporaries, and the live output, but excludes allocations outside PyTorch's CUDA allocator.

Batch	Route	Latency (ms/batch)	Peak memory (MiB)
1	Patch-only	9.610	377.3
1	MHAP + registers	11.577	439.3
8	Patch-only	9.420	477.5
8	MHAP + registers	11.717	541.4

FID compares image distributions in Inception feature space (Heusel et al., 2017). PSNR measures paired pixel fidelity, and DISTS measures paired perceptual differences in structure and texture (Ding et al., 2021). FWD compares distributions of wavelet-packet features (Appendix G.5).

Supplementary FD-DINOv2 measurements use DINOv2 features and are denoted rFDD for reconstruction and gFDD for generation. They are raw distributional distances, distinct from paired feature errors. References and feature-extraction procedures are held fixed within each reported metric comparison.

PSNR aggregation. PSNR is computed separately for each reconstructed image and then averaged over the evaluation set. For RGB values in $[0, 1]$,

$$\overline{\mathrm{PSNR}} = -\frac{10}{n}\sum_{i=1}^{n}\log_{10}\mathrm{MSE}_i\,.$$

This differs from applying the logarithm to arithmetic-mean MSE. We therefore calculate relative reductions in mean pixel error directly from the MSE measurements rather than converting an average PSNR difference into an arithmetic-mean error reduction. The channel-mean and centered contributions are additive in MSE, not in PSNR. DISTS differences are reported in the metric's own units, rather than interpreted as percentages of perceived quality.

A.3 Compute and inference accounting

Register recovery evaluates MHAP once per sampled patch representation, followed by the image decoder; it adds no diffusion sampling pass. MHAP contains 15,366,144 parameters. We measure the patch-only and MHAP-plus-register routes on one NVIDIA A100-PCIE-40GB with BF16 autocast, using the epoch-16 decoders and the MHAP model

At batch sizes 1 and 8, MHAP-plus-register decoding increases latency by 20.47% and 24.38%, and peak allocated memory by 16.42% and 13.38%, respectively. Measurements include input-coordinate conversions and GPU uint8 image conversion. They exclude diffusion sampling, image encoding, model loading, and external input/output transfers. These measurements characterize one model pair; repeated forwards do not estimate variation across trained models.

B Additional Appearance Probes

B.1 Individual-register and register-bank probes

Table 5 compares mean-pooled patches, patches with one register, and patches with the complete register bank. The register identified as `reg4` was selected from the appearance-probe analysis rather than on a separate selection set. Its result therefore describes the selected register, not an unbiased estimate of selecting a register for a new dataset.

The complete bank gives the highest thumbnail and nonzero low-frequency DCT scores. The selected individual register gives higher color-histogram and DC scores than the complete bank. Adding more register features does not therefore improve every linear prediction target.

Table 5: **Individual-register and register-bank probes.** Held-out R^2 from mean-pooled patches, with the selected register, or with all four registers. DC denotes the spatial mean; AC denotes nonzero low-frequency DCT coefficients.

Target	Patch mean	+reg4	+Register bank
Thumbnail	0.211	0.223	0.228
Color histogram	0.288	0.316	0.296
DCT DC	0.735	0.827	0.820
DCT AC	0.038	0.040	0.068

Table 6: **Full-patch appearance probes.** Held-out R^2, averaged across three shared outer splits. Changes are computed before rounding.

Target	Full patches	+Registers	ΔR^2
RGB DC / global mean	0.57336	0.66280	+0.08944
Very-low DCT	0.46391	0.46386	-0.00005
Low DCT	0.16085	0.16082	-0.00003
Mid DCT	0.00466	0.00437	-0.00029
High DCT	-0.00237	-0.00242	-0.00005
Thumbnail	0.24129	0.24711	+0.00582

B.2 Full-patch appearance probes

A separate experiment fits regularized linear probes to the flattened full patch sequence, using online BF16 feature extraction and nested cross-validation. Unlike mean pooling, flattening preserves the positions of all patch features in the probe input. Table 6 reports all evaluated frequency bands and the thumbnail target. Absolute scores are not directly comparable with the patch-mean experiment because the extraction and fitting protocols differ.

The largest improvement is in mean RGB prediction: DC R^2 increases by 0.08944, while thumbnail R^2 increases by 0.00582. Aggregate non-DC R^2 changes by -0.00019.

With identical splits and effective regularization candidates, zero-register inputs reproduce the patch-only predictions exactly. Shuffled-register and Gaussian-register ablation experiments give DC increments of approximately zero and -0.00095, respectively; their thumbnail increments are -0.00083 and -0.00038.

The matched experiment uses 5,000 images and three shared outer splits, with identical inner splits across conditions. Separately row-normalized patch and register features define the additive kernel $K_P + \lambda K_R$, keeping patch-kernel scaling fixed. Regularization and auxiliary-kernel weight are selected within the training data for each target coordinate. Encoder revision, preprocessing, image identities, features, and split assignments are retained.

B.3 Reconstruction from registers alone

The separate DINOv2 ViT-B decoder in Figure 2 receives the encoder's four registers, without spatial patch features or encoder CLS. The four register positions are interpolated to 256 decoder positions to reconstruct a 256×256 image. The decoder is trained on ImageNet-100 with seed 0; the displayed outputs use EMA weights after 15 completed epochs of a 16-epoch schedule. The main-paper curtain, aircraft-wing, terrier, window-screen, and overcast waterfront are selected after inspecting the outputs. Their full image extent and decoded colors are retained. Figure 5 adds six examples from the same register-only decoder.

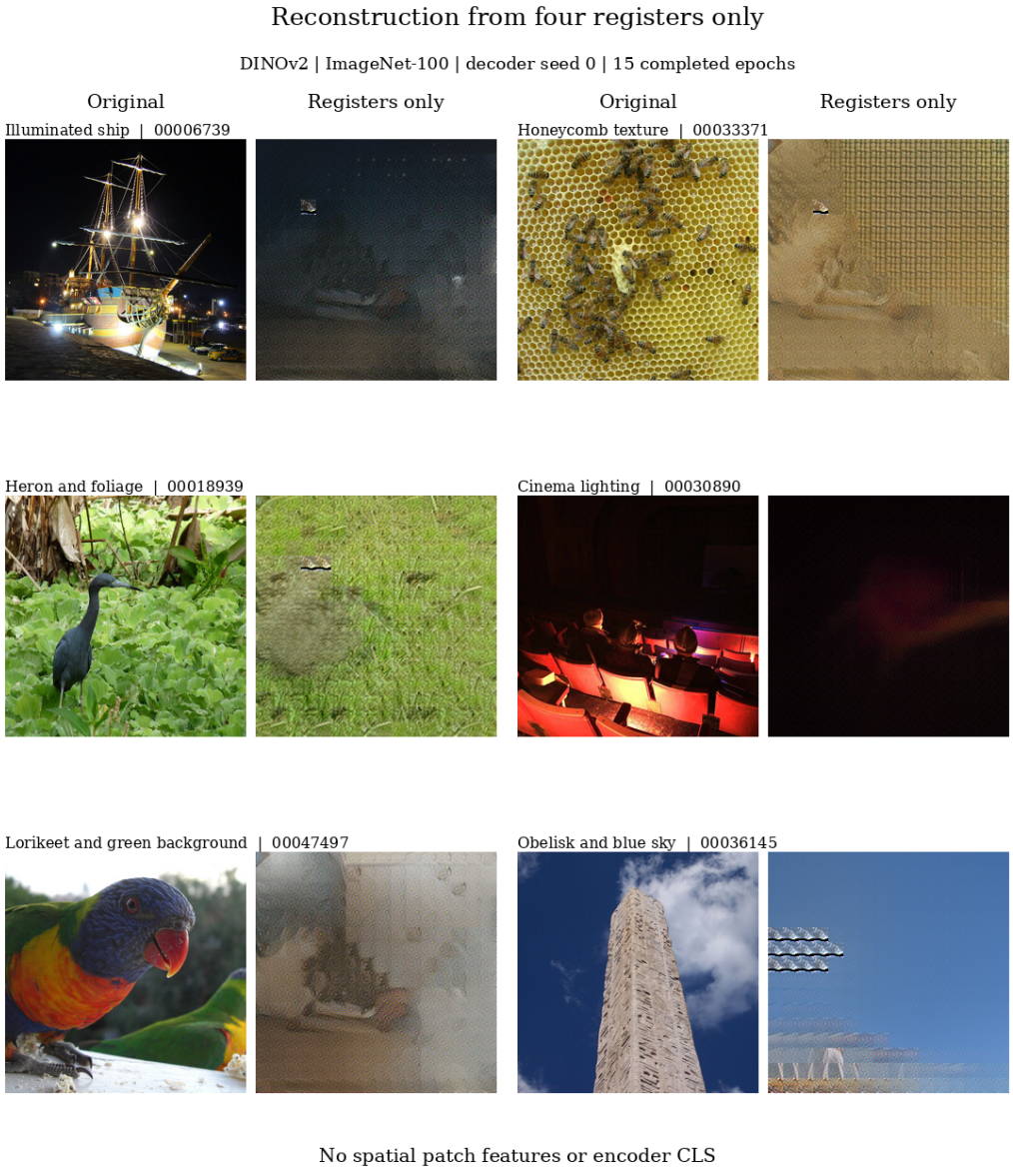

Figure 5: **Additional reconstructions from four registers alone.** Each adjacent pair shows an original photograph and its register-only reconstruction. Examples show an illuminated ship, honeycomb texture, foliage, cinema lighting, plumage, and blue sky. The ImageNet-100-trained DINOv2 ViT-B decoder receives four encoder registers, without spatial patch features or encoder CLS. All outputs use decoder seed 0 and EMA weights after 15 completed epochs.

Table 7: **Non-spatial decoder inputs.** Reconstruction FID in the DINOv2 and DINOv3 input studies. P denotes the complete spatial patch sequence and REG denotes the four encoder registers. Comparisons are made within each encoder's experiment.

Decoder input	DINOv2 rFID ↓	DINOv3 rFID ↓
P	4.8388	4.49223
P + CLS	4.7474	4.37634
P + REG	4.6113	4.33774
P + REG + CLS	4.5932	4.33395

C Decoder Ablations and Reference Results

C.1 Alternative register integration

The principal decoder prepends register features to the patch sequence and processes both through joint self-attention. Two alternative interfaces expose the same type of non-spatial input through modulation or cross-attention.

Pooled modulation compresses the register bank into a vector $q = \mathrm{Pool}(Z_r)$ and predicts a feature-wise affine transformation:

$$(\gamma, \beta) = g(q), \qquad \widetilde{h}_\ell = \gamma \odot h_\ell + \beta. \tag{14}$$

Cross-attention instead uses projected register features as a memory, with queries from the patch stream.

These interfaces differ in how register information enters the decoder. Modulation conditions patch features through a pooled vector; cross-attention reads a separate register memory; prepending allows patches and registers to evolve jointly. Results involving CLS-plus-register inputs are identified separately from the four-register-only comparison.

C.2 Choice of non-spatial decoder inputs

We compare decoding from patches alone with the addition of CLS, four registers, or their combination. All conditions retain the complete spatial patch sequence. Table 7 reports the reconstruction FID for each encoder's input study.

Register inputs yield lower rFID than CLS inputs in both studies. Adding CLS to the register bank reduces rFID by a further 0.0181 for DINOv2 and 0.00379 for DINOv3, corresponding to 0.39% and 0.09% of the respective register-input scores. This pattern is consistent with the appearance-related specialization of register features described by Fel et al. (2026) and motivates the four-register interface used in our principal experiments. The decoder-capacity study below evaluates the combined CLS-plus-register input and retains that input definition.

C.3 Decoder capacity

The capacity experiment asks whether a larger patch-only decoder removes the advantage of supplying non-spatial encoder outputs. The ImageNet-100 comparison evaluates ViT-B, ViT-L, and ViT-XL decoders with DINOv2 features. The augmented models receive the encoder CLS token together with four registers, in addition to all spatial patches. This is a five-token auxiliary-input comparison, not a four-register-only ablation.

The auxiliary-input decoder has lower rFID at each tested size. Increasing capacity from ViT-B to ViT-L improves the patch-only reference, but the auxiliary-input advantage remains. ViT-XL does not further improve the recorded ViT-L scores. The advantage persists across the tested capacity range: 6.2%, 5.5%, and 5.5% for ViT-B, ViT-L, and ViT-XL, respectively.

The longer ViT-L comparison also improves paired fidelity: PSNR increases from 18.9101 to 19.2504 dB and SSIM from 0.4706 to 0.4790. Its clean-cache rFDD decreases from 37.4515 to 35.9324. These measurements use the combined CLS-plus-register input. Four-register-only capacity effects and training-seed uncertainty are unavailable for the B/L/XL sweep.

Table 8: **Decoder capacity with DINOv2 features on ImageNet-100.** Reconstruction FID for patches alone and for patches with CLS and four registers. Epochs denote the recorded training-schedule length. The 16-epoch B/L/XL values are recorded manuscript measurements without retained seed-level metric records. The 40-epoch ViT-L row is a separately retained seed-0 evaluation. Relative reductions are computed within each row.

Decoder	Epochs	Patches	+CLS+4REG	Reduction
ViT-B	16	5.7289	5.3722	6.2%
ViT-L	16	5.4250	5.1241	5.5%
ViT-XL	16	5.4720	5.1724	5.5%
ViT-L	40	4.6091	4.4234	4.0%

Table 9: **Register supervision and register inputs.** DINOv2 ImageNet-1K, decoder seed 0, 16 epochs. Parameter counts are in millions for reconstruction/generation, excluding the frozen encoder, diffusion generator, and training-only prediction head. Counts include fixed positional parameters and, for InfoNCE, 3,072 stored null-slot parameters unused during ordinary inference.

Approach	Parameters (r/g)	rFID ↓	gFID ↓
Patch-only	86.435 / 86.435	0.5253	1.8189
Internal InfoNCE	143.806 / 143.806	0.5104	1.7888
Internal register MSE	88.215 / 88.215	0.5181	1.8399
Native / MHAP registers	87.029 / 102.395	**0.4708**	**1.6561**

C.4 Register supervision versus register inputs

We compare two ways of using four terminal encoder registers as training targets with supplying registers through the decoder interface. All rows use DINOv2 ViT-B features and decoder seed 0 at the 16-epoch ImageNet-1K endpoint, with $\tau = 0$. Reconstruction uses the same 50,000 validation images. Generation uses the same retained 50,000 pretrained-XL patch samples and evaluation reference.

Internal register supervision. The contrastive variant uses four learned slots with bidirectional cross-attention between patches and slots. After two completed decoder blocks, a LayerNorm–linear head predicts the four standardized encoder registers. Its additional objective is

$$0.05\,\mathcal{L}_{\mathrm{InfoNCE}} + 0.05\,\mathcal{L}_{\mathrm{slot}}.$$

InfoNCE compares corresponding register slots across images, with temperature 0.07. The slot regularizer combines a variance hinge with target standard deviation 0.5 and the mean squared cosine similarity between distinct slots. Slot dropout is 0.1 during training. The prediction head is removed for inference; the learned slots and cross-attention remain. Patches are the only image-dependent inference input.

The register-MSE variant instead prepends four learned slots. After two completed self-attention blocks, a LayerNorm–linear head predicts four standardized registers. A learned linear projection maps these predictions back to decoder width and replaces the four slot states before the remaining ten blocks. The register MSE has coefficient 0.1. The prediction head and projection remain in the image path, so the image objective also trains this prediction pathway. This variant uses neither native registers nor an external predictor at inference, and uses no slot dropout.

The contrastive variant improves both recorded FIDs relative to the patch-only reference. The register-MSE variant improves reconstruction FID but increases generation FID. Neither reaches the scores obtained with native registers for reconstruction and MHAP predictions for generation. These comparisons show that the tested uses of register supervision and register inputs produce different outcomes. Because decoder architectures and objectives also differ, they do not isolate the effect of loss choice or establish a parameter-matched comparison.

Table 10: **Decoder-context ablations on ImageNet-100.** Single-seed measurements with five added tokens. Neither context condition receives non-spatial encoder outputs. The ablation decoders are not token-count matched to the four-register decoder.

Decoder input	rFID ↓	rFDD ↓
Patches	4.8284	41.3454
Patches + 5 learned tokens	4.7992	41.3672
Patches + 5 PATCHMAP summaries	4.8168	40.9061

Table 11: **ImageNet-100 decoder-input comparison.** Mean and reported standard deviation across three decoder seeds. Both routes receive all spatial patch features; the second also receives the encoder's four register outputs.

Decoder input	rFID ↓
Patches	4.8388 ± 0.0302
Patches + registers	$\mathbf{4.6113} \pm 0.0087$

C.5 LEARNED DECODER CONTEXT AND PATCH SUMMARIES

The decoder-context ablations add five tokens without supplying encoder CLS or register outputs. One condition uses learned image-independent tokens. The other produces five PATCHMAP summary tokens from the spatial patch sequence. Table 10 compares both with their own seed-0 patch-only reference.

Learned tokens reduce rFID by 0.0292 while increasing rFDD by 0.0218. Patch summaries reduce rFID by 0.0116 and rFDD by 0.4393. These ablations use one seed and five tokens; the principal register result uses four tokens and three decoder seeds.

D SUPPLEMENTARY RECONSTRUCTION MEASUREMENTS

D.1 IMAGENET-100 DECODER-INPUT COMPARISON

The decoder-input study compares patches alone with patches and four registers across three decoder-training seeds. All spatial patches are retained and both conditions use the same base image objective. These models are separate from the A/B appearance-analysis models in Section 4.2.

D.2 IMAGENET-1K RECONSTRUCTION BY DECODER SEED

Table 12 supplies the per-seed measurements for Table 2. Each reconstruction set contains the same 50,000 validation image identities within an encoder comparison. Means and sample standard deviations are computed across the three decoder seeds. The aggregate rows reproduce the principal reconstruction comparison in Table 2.

D.3 IMAGENET-1K REGISTER-DECODER FIDELITY

Table 13 reports paired image fidelity and wavelet distance for the register-input decoders.

DINOv3's higher PSNR accompanies lower mean RGB MSE, 0.01134 versus 0.01334 for DINOv2. Its lower DISTS indicates closer agreement with the source images under that structure-and-texture metric.

E OUT-OF-DISTRIBUTION RECONSTRUCTION

E.1 IMAGENET-1K-TRAINED DECODERS

We evaluate ImageNet-1K-trained DINOv2 and DINOv3 decoders on NINCO and Places365 without additional training (Bitterwolf et al., 2023; Zhou et al., 2018). The evaluation sets contain 5,878

Table 12: **ImageNet-1K reconstruction by decoder seed.** All decoders use the ViT-B architecture. Register-input decoding supplies four encoder registers and the complete patch sequence. The aggregate rows reproduce the principal reconstruction comparison.

Encoder	Decoder seed	Patches rFID ↓	+Registers rFID ↓
DINOv2	0	0.525288	0.470804
	1	0.516679	0.469105
	2	0.521390	0.475329
	Mean ± SD	0.521119 ± 0.004311	0.471746 ± 0.003217
DINOv3	0	0.464152	0.443066
	1	0.467069	0.440756
	2	0.469339	0.443407
	Mean ± SD	0.466853 ± 0.002600	0.442410 ± 0.001442

Table 13: **Fidelity of ImageNet-1K register-input reconstructions.** Mean ± sample standard deviation across three decoder seeds. PSNR and DISTS compare each reconstruction with its source image; FWD compares image distributions in wavelet coordinates. Both rows use patches and four encoder registers.

Encoder	PSNR (dB) ↑	DISTS ↓	FWD ↓
DINOv2	19.7262 ± 0.0012	0.10444 ± 0.00005	0.4579 ± 0.0105
DINOv3	20.5406 ± 0.0156	0.09670 ± 0.00007	0.4443 ± 0.0050

NINCO images and 36,500 Places365 images. Within each encoder comparison, both routes use a ViT-B decoder backbone and the 16-epoch endpoint. We compare decoding from the complete patch sequence with decoding from patches and four native encoder registers. Evaluation images, preprocessing, output conversion, and metric implementations are held fixed across the two routes. Table 14 reports reconstruction FID and paired image fidelity.

Register-input decoding improves all three metrics for both encoder families on both datasets. For DINOv2, rFID decreases by 6.18% on NINCO and 43.93% on Places365. The corresponding reductions for DINOv3 are 5.18% and 28.14%. All percentages are computed from unrounded scores.

Paired reconstruction fidelity improves alongside rFID. On NINCO and Places365, respectively, DINOv2 gains 0.3089 and 0.1686 dB in PSNR, while DISTS decreases by 0.00652 and 0.00526. DINOv3 gains 0.4022 and 0.1863 dB in PSNR, with DISTS reductions of 0.00588 and 0.00489. These results show that the reconstruction benefits extend to images outside the decoder-training dataset.

Among register-input decoders, DINOv3 achieves higher PSNR and lower DISTS on both datasets. The rFID ordering depends on the dataset: DINOv3 scores lower on NINCO, whereas DINOv2 scores lower on Places365.

E.2 ImageNet-100-trained patch-only decoders

Table 15 reports patch-only references for decoders trained on ImageNet-100. These differ in training dataset from the ImageNet-1K register decoders above and are not paired counterparts to Table 14.

F Register Recovery and Generation Details

F.1 Image-space evaluation of recovered registers

Before applying register recovery to generated patches, we evaluate its image-space consequences on 50,000 ImageNet-1K validation images. We hold one DINOv2 register-input decoder and each image's complete patch sequence fixed, and compare native encoder registers, MHAP predictions, training-set mean registers, and shuffled MHAP predictions (Table 16).

Table 14: **Reconstruction outside ImageNet-1K.** ImageNet-1K-trained ViT-B decoders evaluated after 16 epochs, without training on either evaluation dataset. All results use decoder seed 0.

Encoder	Decoder input	rFID ↓	PSNR (dB) ↑	DISTS ↓
NINCO				
DINOv2	Patches	4.9567	19.9535	0.12304
	Patches + registers	4.6502	20.2624	0.11652
DINOv3	Patches	4.6931	20.7924	0.11497
	Patches + registers	4.4501	21.1945	0.10909
Places365				
DINOv2	Patches	2.5935	19.1880	0.11521
	Patches + registers	1.4542	19.3566	0.10995
DINOv3	Patches	2.5663	20.0437	0.10823
	Patches + registers	1.8441	20.2301	0.10334

Table 15: **ImageNet-100-trained patch-only OOD references.** DINOv2 patch-only decoders trained on ImageNet-100. The table preserves the reported reference scores without adding unmeasured register-decoding comparisons.

Evaluation dataset	rFID ↓	rFDD ↓
ImageNet-R	2.44	22.07
NINCO	6.62	65.76
Places365	4.88	26.30

Predictions give lower pixel and perceptual errors than mean or shuffled replacements, showing the value of receiver-specific recovery. Encoder registers remain better: substituting predictions increases DISTS by 0.00764 (paired image-bootstrap 95% interval $[0.00756, 0.00772]$, conditional on the fixed models). Mean-color error accounts for most of the total MSE increase. High feature agreement therefore does not imply equivalent decoded images (Appendix F.3).

F.2 Register replacements on generated patches

We test register-input specificity using the existing 50,000 pretrained DiT-DH/XL patch samples. The DINOv2 four-register decoder, MHAP predictor, sampled patches, token positions, output processing, and official ADM ImageNet-256 population reference remain fixed. We compare aligned MHAP predictions with the per-slot training-set mean and with within-class shuffled predictions. Shuffling permutes complete four-register banks while preserving slot order; permutation seed 0 produces zero self-assignments and zero cross-class assignments. MHAP is evaluated before replacement in both control routes.

Aligned predictions yield lower FID than either replacement, supporting the usefulness of sample-specific register values in the trained decoder.

F.3 MHAP predictor

Architecture and inputs. We fit a separate predictor for each encoder, using its complete sequence of 256 terminal patch features of dimension 768 to predict four register features in encoder-defined order. A learned 768×768 projection and LayerNorm transform the standardized patch features. Four learned 768-dimensional queries pass through two attention-pooling blocks, each with 12-head cross-attention over the same patch sequence and a $768 \to 3072 \to 768$ GELU feed-forward network. Attention and feed-forward updates have residual connections and pre-LayerNorm; queries and source features have separate attention-input LayerNorms. A final LayerNorm and linear projection produce four 768-dimensional outputs. No additional positional embeddings, query self-attention, dropout, or feature noise are used. The predictor has 15,366,144 parameters and does not receive or predict CLS.

The frozen teacher uses terminal features with the final LayerNorm affine parameters removed. Each source photograph receives the fixed 256-pixel center crop; DINOv2 inputs are then resized to

Table 16: **Register recovery with a fixed decoder.** One decoder–predictor pair, 50,000 ImageNet-1K images, and identical receiver patches. The MSE columns correspond to the two terms in Equation 5; lower values indicate smaller errors.

Register source	Mean-color MSE ($\times 10^{-3}$)	Centered MSE ($\times 10^{-3}$)	DISTS ↓
Encoder registers	0.5665	12.7735	0.10441
MHAP predictions	1.6368	13.0146	0.11205
Training-set mean	4.5069	13.5449	0.12890
Shuffled MHAP	6.2529	13.8718	0.13331

Table 17: **Register replacements during generation.** One fixed decoder–predictor pair and identical sampled patches. Relative changes use the aligned seed-0 score as denominator.

Register source	gFID ↓	Change from aligned
Aligned MHAP	**1.6561**	—
Within-class shuffled MHAP	1.7981	+8.57%
Training-set mean	2.3155	+39.81%

224×224 by bicubic interpolation, whereas DINOv3 uses 256×256. Both use the pinned encoder processor's RGB mean and standard deviation. Teacher features are extracted online in microbatches of 16 and retained at a BF16 feature boundary. Predictor training is independent of the image decoder and diffusion model.

Standardization and objective. Statistics are fitted separately for each encoder using only the 1,279,690-image scientific training subset. Patch standardization uses one mean and standard deviation per feature channel, pooled across images and all 256 patch positions. In particular, if μ_{cs} and v_{cs} are the population moments at channel c and position s,

$$\bar{\mu}_c = \frac{1}{256} \sum_s \mu_{cs}, \qquad \bar{v}_c = \frac{1}{256} \sum_s (v_{cs} + \mu_{cs}^2) - \bar{\mu}_c^2.$$

This includes between-position variation. Register targets use separate population moments for each of the 4×768 slot–channel coordinates, pooled over images. In both cases the divisor is $\max(\sqrt{\max(v, 0)}, 10^{-6})$: the predictor floors the standard deviation rather than adding an epsilon inside the square root. Population moments use FP64 accumulation and are saved in FP32.

Let $\mathcal{S}_p$ and $\mathcal{S}_r$ denote these patch and register standardizations, and let g_ϕ return standardized register predictions. Define the encoder-coordinate predictor by $f_\phi = \mathcal{S}_r^{-1} \circ g_\phi \circ \mathcal{S}_p$. We minimize

$$\mathcal{L}_{\text{reg}}(\phi) = \mathbb{E}_{x \sim \mathcal{D}_{\text{train}}} \left[\text{MSE}(\mathcal{S}_r(f_\phi(Z_p(x))), \mathcal{S}_r(Z_r(x))) \right]. \tag{15}$$

MSE is the unweighted mean over the minibatch, four registers, and 768 feature dimensions. Loss arithmetic is FP32; teacher and predictor forwards use BF16, with TF32 disabled for the teacher and enabled for the predictor.

Training and selection. Both encoders use the same optimization settings: AdamW with learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, and batch size 256, retaining the final partial batch. We train for a fixed 100 epochs (499,900 updates), with five epochs of linear warmup followed by cosine decay to 10^{-5} and gradient-norm clipping at 1.0. The 640 reserved calibration parents and content duplicates are excluded from predictor fitting and normalization-statistic estimation. Checkpoint selection minimizes standardized register MSE on 128 fixed, training-derived selection images after each epoch; exact ties retain the earlier epoch. Another 256 disjoint training-derived images provide monitoring only. Final validation images are excluded from fitting and checkpoint selection. The seed-0 checkpoints are selected at epoch 96 for DINOv2 and epoch 95 for DINOv3. All three DINOv2 decoder-seed evaluations reuse the epoch-96 predictor and the same generated-patch bank.

Generation coordinate conversion. For the DINOv2 pretrained-XL experiment, source normalization of each stored generated patch tensor is inverted exactly once. Raw patch features are

Table 18: **DINOv2 generation by decoder seed.** All rows decode the same 50,000 pretrained-XL patch samples. MHAP uses predictor seed 0, selected at epoch 96, for every register decoder. Aggregate rows report mean ± sample standard deviation across decoder seeds. The released decoder is a separate ViT-XL reference.

Decoder seed	Patch-only ViT-B gFID ↓	MHAP + REG4 ViT-B gFID ↓
0	1.818943	1.656117
1	1.813372	1.637629
2	1.824930	1.630376
Mean ± SD	1.819082 ± 0.005780	1.641374 ± 0.013273
Released ViT-XL decoder		1.557247

standardized for MHAP and independently converted to the receiving RAE wrapper's patch coordinates. MHAP outputs return to encoder coordinates through $\hat{r}_{jc} = \mu_{r,jc} + \sigma_{r,jc}\hat{r}^{\mathrm{std}}_{jc}$. The generation adapter then applies the receiver's auxiliary normalization and calls `decode` with `aux_tokens_are_normalized=True`. The RAE wrapper *inverts* that normalization before passing the registers to the learned decoder's projection and prepend interface. Native reconstruction instead supplies raw register features with this flag set to `False`. Thus both routes supply encoder-coordinate register features to the learned decoder; generation's extra normalize/inverse-normalize pair is an API coordinate conversion. These wrapper transforms use $\sqrt{v + 10^{-5}}$, distinct from the predictor's standard-deviation floor. The wrapper likewise inverts its patch normalization before decoding. No image, native-register target, or CLS feature is available to the generation path.

F.4 DINOv2 pretrained-XL evaluation

The primary generation comparison uses a fixed bank of 50,000 DINOv2 patch representations sampled by the pretrained DiT-DH/XL. The bank was generated using the Euler ODE sampler with `num_steps=50`, classifier-free guidance scale 1.0, and global sampling seed 42. Sampling uses BF16 with TF32 enabled. Every decoding route preserves the same stored patch tensors, image identities, and class labels.

The stored patch bank uses FP16 tensors. Source normalization is inverted once before conversion to the receiving decoder's coordinates. The evaluated decoding procedure uses BF16 with TF32 enabled and microbatches of eight. Outputs are clamped to $[0, 1]$, multiplied by 255, and truncated to uint8 before evaluation. All routes use the same pinned Inception evaluator and real-image reference.

The released RAE decoder has gFWD 0.71607 against a fixed set of 10,000 real RGB images. This FWD reference differs from the ADM-reference evaluation used for gFID. The two measurements therefore use different representations and reference supports. Reconstruction FWD uses its own evaluation cohort and is not numerically interchangeable with generation FWD.

All decoding routes receive the same 50,000 stored patch samples. Their outputs are evaluated using identical image postprocessing, the same Inception evaluator, and the same real-image reference. Training details for the register-input decoders appear in Appendix A.1. The released ViT-XL decoder is larger than the ViT-B register decoder and uses its own training recipe.

F.5 Joint generation of patches and auxiliary tokens

We also examine a separate ImageNet-1K experiment that jointly generates DINOv2 patches and five auxiliary tokens: CLS and four registers. This experiment uses DiT-S generators trained for 80 epochs, rather than the pretrained XL generator used in the main comparison. The generators use a 12-block, width-384 backbone with six attention heads and a two-block, width-2048 DDT head with 16 attention heads.

Joint latent modeling. The joint generator uses separate timestep and token-type embeddings for the two streams. Patches follow a linear transport path, while auxiliary tokens use a GVP path. We compare auxiliary-loss weights $\lambda_{\mathrm{aux}} \in \{0.05, 0.25, 0.50\}$. The patch-only and joint generators use

Table 19: **Joint auxiliary generation.** Separate DiT-S experiment with five auxiliary tokens (CLS and four registers), $\tau = 0.8$ receivers, and 50,000 evaluated images. † The retained metric row does not identify its FID reference; this value is not used for cross-route comparisons.

Generation and decoding route	gFID ↓
Patch generation, patch-only decoding	5.5981
Patch generation, MHAP auxiliary recovery	**5.4041**
Joint generation, $\lambda_{\text{aux}} = 0.05$	7.7431
Joint generation, $\lambda_{\text{aux}} = 0.25$	6.8716
Joint generation, $\lambda_{\text{aux}} = 0.50$	$7.1076^{\dagger}$

the same 80-epoch training budget, effective batch size 1024 with four-step gradient accumulation, and training seed 42. Their optimizer is AdamW with initial learning rate 2×10^{-4}, $\beta = (0.9, 0.95)$, and zero weight decay. The learning rate remains constant for the first 40 epochs and then decays linearly toward 2×10^{-5} at a configured 800-epoch endpoint; training stops at epoch 80.

Sampling and decoding. Each evaluation uses 50,000 class-balanced samples, 50-step Euler integration, CFG scale 1.0, sampling seed 42, and EMA generator weights. The ViT-B receivers are decoder-seed-0 models trained for 16 epochs with patch-latent noise $\tau = 0.8$. The patch-only receiver consumes patches; the auxiliary receiver prepends CLS and four registers. For auxiliary recovery, a separately trained MHAP predictor uses five learned queries, two attention-pooling blocks, hidden width 768, 12 heads, and MLP ratio 4. It predicts CLS and four registers from generated patches; it is distinct from the four-register predictor used with the pretrained-XL samples. The five-token predictor uses seed 0 and the retained best checkpoint from a 100-epoch training run.

These scores belong to the separate DiT-S experiment and are not directly comparable with the pretrained-XL evaluation. The joint-generation rows at auxiliary-loss weights 0.05 and 0.25 record the ImageNet-1K validation reference; the reference identity for weight 0.50 is unavailable.

At auxiliary-loss weights 0.05 and 0.25, joint generation has higher gFID than patch generation followed by MHAP recovery under the recorded common reference. This compares separately trained generators, not different auxiliary sources at fixed sampled patches. The experiment concerns CLS plus four registers; four-register-only joint generation and the causal effect of the transport-path choice are not isolated.

F.6 Qualitative model and sample identities

Composite reconstruction and generation examples. Figure 4(a) shows reconstruction evaluation IDs 026865 (dogsled), 012929 (Samoyed), and 048933 (seashore), from top to bottom. Panel (b) shows generated sample IDs 005256 (night stage), 001171 (underwater shark), and 001530 (heron on water).

All displayed decoder outputs use DINOv2 features and seed-0 ViT-B decoders trained on ImageNet-1K, evaluated after 16 epochs. The patch-only model is the reference checkpoint. Reconstruction supplies four native encoder registers to the register-input decoder; generation supplies four predictions from the seed-0 MHAP checkpoint selected at epoch 96.

Reconstruction triplets share the same evaluated source crop. Each generation pair uses the same stored pretrained-DiT-DH/XL patch tensor and has no paired source image. The reconstruction examples were selected post hoc from the retained candidate gallery to illustrate appearance differences; they do not estimate the average improvement. Generation examples were drawn from the existing retained gallery. Quantitative comparisons use their complete fixed evaluation sets. All photographs retain their complete 256×256 pixels, without additional cropping, color adjustment, sharpening, or other enhancement.

Table 20: **DISTS component changes on ImageNet-100.** Values average the three decoder-seed means. Changes are register-input minus patch-only. Confidence intervals use paired seed differences and Student's t with two degrees of freedom.

Contribution	Patch-only	Registers	Change	95% interval
Texture	0.085537	0.079748	-0.005789	$[-0.006686, -0.004891]$
Structure	0.037023	0.035759	-0.001263	$[-0.001549, -0.000977]$
Total DISTS	0.122559	0.115507	-0.007052	$[-0.008227, -0.005877]$

G Appearance Analysis and Evaluation

G.1 DISTS texture and structure contributions

We decompose DISTS for the ImageNet-100 appearance experiment. Each of three decoder seeds reconstructs the same 5,000 validation images using patch-only (A) and register-input (B) decoding. This gives 30,000 source–reconstruction evaluations and 15,000 paired A/B image–seed comparisons.

Let $S_{\mu,j}$ and $S_{\mathrm{cov},j}$ denote the implemented feature-mean and feature-covariance similarities, respectively. The weighted texture and structure contributions are

$$\begin{aligned} W &= \sum_j (\alpha_j + \beta_j), \\ T &= \sum_j \frac{\alpha_j}{W}(1 - S_{\mu,j}), \qquad S = \sum_j \frac{\beta_j}{W}(1 - S_{\mathrm{cov},j}). \end{aligned} \tag{16}$$

The sums include RGB and five frozen modified-VGG feature banks, following DISTS (Ding et al., 2021). Both contributions retain the same normalization W; neither is divided by its own component weight mass.

Both contributions decrease in every decoder seed. Texture and structure account for 82.1% and 17.9% of the weighted DISTS reduction, respectively. These percentages partition the metric change, not the amount of preserved texture or the causes of the reconstruction improvement. The maximum per-image discrepancy between DISTS and $T + S$ is 1.975×10^{-7}, and all 30,000 total scores reproduce the previously recorded DISTS values.

G.2 Residual spectra

The Fourier analysis uses 30,000 A/B reconstructions: 5,000 images, two decoding routes, and three decoder seeds. Smooth and detailed windows are selected from the original images and used for both reconstructions.

Relative to patch-only decoding, register-input decoding reduces residual-gradient MSE by 1.372×10^{-5} in smooth windows, with a 95% seed interval of $[1.120, 1.624] \times 10^{-5}$. In detailed windows, the reduction is 9.117×10^{-4}, with interval $[7.538, 10.696] \times 10^{-4}$.

Output gradient power also decreases in detailed windows. Thus, lower residual-gradient error co-occurs with lower output gradient power in detailed regions. The measurements describe changes in fidelity to the original gradient field alongside changes in the reconstructed field itself. Radial averaging removes orientation information; this analysis measures residual magnitude by radial frequency.

G.3 Register substitutions

These measurements use the ImageNet-100 `B_preg_standard` decoders, training seeds 0, 1, and 2, on the same 5,000 validation images. They are distinct from the ImageNet-1K experiments. The four-way comparison in Equation 6 crosses original and transformed patch/register states within each fixed decoder; decoder weights, register positions, normalization, and output procedure remain fixed.

Joint color and saturation transformation. For each spatial location of the original sRGB-coded image $x \in [0,1]^{3\times H\times W}$, define

$$g = (1.08, 0.98, 0.90), \quad b = (0.015, 0, -0.005), \tag{17}$$

$$y = g \odot x + b, \ell \qquad = 0.2126y_R + 0.7152y_G + 0.0722y_B, \tag{18}$$

$$T(x)_c = \mathrm{clip}_{[0,1]}\big(\ell + 1.12(y_c - \ell)\big). \tag{19}$$

Thus the saturation pivot is the per-pixel luma of the *already gain-and-bias-transformed* RGB vector. There is no separate luminance-contrast pivot, spatial resampling, or conversion to linear-light RGB. Clipping occurs only after the saturation operation; 4.1886% of the channel values require clipping. The transformed image remains floating point for encoding and readout. All intervention metrics use decoder outputs clamped to $[0,1]$ *before* rounding to the archived uint8 reconstructions.

Directional aggregation. Let $m(z) \in \mathbb{R}^3$ be the vector of spatial channel means, $d_i = m(T(x_i)) - m(x_i)$, and

$$u^P_{s,i} = m(D_s(P_i, R_{T,i})) - m(D_s(P_i, R_i)), \tag{20}$$

$$u^{P_T}_{s,i} = m(D_s(P_{T,i}, R_{T,i})) - m(D_s(P_{T,i}, R_i)). \tag{21}$$

For either patch state $Q \in \{P, P_T\}$, the within-seed coefficient is

$$\beta^Q_s = \frac{\sum_{i\in\mathcal{I}} (u^Q_{s,i})^\top d_i}{\sum_{i\in\mathcal{I}} \|d_i\|_2^2}, \qquad \mathcal{I} = \{i : \|d_i\|_2 \geq 10^{-3}\}. \tag{22}$$

This is a pooled slope through the origin, equivalently a $\|d_i\|_2^2$-weighted mean of the individual projection coefficients. It is not the unweighted mean of per-image ratios. Below the threshold, projection, cosine, numerator, and denominator readouts are marked undefined and excluded from pooling; no additive denominator epsilon defines a finite slope there. All 5,000 RGB directions are defined in each of the four cells for all three seeds.

Reported coefficients are the arithmetic means of the three within-seed slopes. The original-patch and transformed-patch coefficients are respectively 0.49943 and 0.55917; the paired interaction is 0.05974. The four cells share the same directions, denominator, and support within each seed, so the transformed-patch effect is the difference of their pooled slopes. Decoder training seed is the outer unit: 95% intervals use $\bar\beta \pm 4.30265\, s_\beta/\sqrt{3}$, with sample standard deviation s_β across seeds.

Orthogonal response. For the register substitution at *original* receiver patches, the per-image orthogonal response is

$$o_{s,i} = \left\| u^P_{s,i} - \frac{(u^P_{s,i})^\top d_i}{\|d_i\|_2^2} d_i \right\|_2 . \tag{23}$$

Its reported value, 0.01552 in $[0,1]$ RGB units, is the arithmetic mean of these Euclidean norms over images within seed and then over three seeds. It is not the norm of an averaged response or an RGB RMS. For a direction below threshold, the implementation retains the full change norm as a finite diagnostic by setting the subtracted component to zero; no such cases occur in this appearance comparison.

Edge and semantic readouts. The edge statistic is cosine similarity between flattened *Sobel-magnitude* maps. For the sRGB-coded grayscale combination $L(z) = 0.2126z_R + 0.7152z_G + 0.0722z_B$, define

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \qquad K_y = K_x^\top, \tag{24}$$

$$e(z) = \mathrm{vec}\,\sqrt{(K_x * L(z))^2 + (K_y * L(z))^2 + 10^{-12}}. \tag{25}$$

Convolutions use one-pixel zero padding and the unnormalized kernels shown above. Cosine similarity clamps each vector norm below at 10^{-8}. The reported change averages

$$\cos(e(D_s(P_i, R_{T,i})), e(x_i)) - \cos(e(D_s(P_i, R_i)), e(x_i))$$

over images within seed and then over seeds, giving $+0.000374$. This compares gradient magnitudes, not gradient orientations.

Under the same original-patch register substitution, supervised ConvNeXt-Tiny logit cosine to the original image changes from 0.92937 to 0.92420. Top-1 agreement changes by -0.0133 percentage points; its 95% seed interval crosses zero. The evaluator uses the fixed ImageNet-1K weights, bilinear $256 \to 236$ resizing, a centered 224×224 crop, and ImageNet channel normalization. These results describe appearance responses and accompanying fidelity/semantic readouts. Channel-wise color and saturation change jointly in this intervention. Independent brightness, color-balance, and contrast control requires separate interventions.

G.4 Register interpolation with fixed receiver patches

Decoder and image pairs. The trajectories in Figure 3 use the DINOv2 four-register ViT-B decoder trained on ImageNet-1K for 16 epochs, decoder seed 0. Each source image is encoded once. During interpolation, receiver patches, decoder weights, and token positions remain fixed; only the four register inputs change. Neither encoder CLS nor MHAP predictions are supplied.

The displayed pairs are `p02_ba` and `p10_ab`. For `p02_ba`, the receiver is `Places365_val_00013976.jpg` and the donor is `Places365_val_00030622.jpg`. For `p10_ab`, the receiver is `ILSVRC2012_val_00031709.JPEG` and the donor is `ILSVRC2012_val_00006179.JPEG`. These display pairs were selected after inspecting the saved trajectories. The source-image experiment contains ten preselected pairs evaluated in both directions, giving 20 retained trajectories.

Slotwise interpolation. For register slot $j \in \{1, \dots, 4\}$, let r_j and $r_j^\star$ be the receiver and donor features in terminal encoder coordinates. For nonzero endpoints, define

$$u_j = \frac{r_j}{\|r_j\|_2}, \qquad v_j = \frac{r_j^\star}{\|r_j^\star\|_2}, \qquad \theta_j = \arccos(u_j^\top v_j).$$

For $0 < \theta_j < \pi$,

$$\begin{aligned} r_j(\alpha) &= \rho_j(\alpha) \left[\frac{\sin((1-\alpha)\theta_j)}{\sin\theta_j} u_j + \frac{\sin(\alpha\theta_j)}{\sin\theta_j} v_j \right], \\ \rho_j(\alpha) &= (1-\alpha)\|r_j\|_2 + \alpha\|r_j^\star\|_2. \end{aligned} \tag{26}$$

Directions follow the shortest spherical arc and norms are interpolated linearly. All four slots use the same α, without slot mixing, centering, or whitening. The eight slot pairs shown in Figure 3 use this non-degenerate spherical branch. At $\alpha = 0$ and $\alpha = 1$, the original endpoint tensors are supplied directly.

We evaluate $\alpha \in \{0, 0.125, \dots, 1\}$. Interpolation arithmetic uses FP64; intermediate registers are converted to FP32 before decoding. Inference uses CPU BF16 autocast with batch size one. Each curve contains the nine recorded measurements for one pair, without averaging across receivers. The image panels display $\alpha = 0, 0.25, 0.5, 0.75, 1$.

Luminance and gradient-energy measurements. Let $x_\alpha = \mathrm{clip}_{[0,1]}(\hat{x}_\alpha)$ denote the decoded RGB image, with $H = W = 256$. Measurements use these clipped floating-point outputs before rounding to uint8 for display. Let $w = (0.2126, 0.7152, 0.0722)^\top$ and define the channelwise sRGB-to-linear conversion

$$\ell(s) = \begin{cases} s/12.92, & s \le 0.04045, \\ ((s + 0.055)/1.055)^{2.4}, & s > 0.04045. \end{cases}$$

Mean linear-light luminance is

$$\overline{Y}(x_\alpha) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{k=1}^{W} w^\top \ell(x_{\alpha,hk}). \tag{27}$$

Table 21: **All retained register-interpolation trajectories.** DINOv2 four-register decoder trained on ImageNet-1K, decoder seed 0. Donor indicates a lower or higher native reconstructed statistic. Aligned counts donor-directed endpoint changes; monotonic counts trajectories following that direction across all eight adjacent steps. Medians include adverse responses.

Statistic	Donor	n	Aligned	Monotonic	Median Δ
Luminance	Lower	10	10	7	-0.092725
Luminance	Higher	10	9	7	+0.049684
Normalized gradient energy	Lower	10	10	8	-0.203004
Normalized gradient energy	Higher	10	10	9	+0.065271

Gradient energy instead uses the sRGB-coded luma $q_{hk} = w^\top x_{\alpha,hk}$. With spatial mean $\bar{q}$, define

$$E(q) = \frac{1}{H(W-1)} \sum_{h=1}^{H} \sum_{k=1}^{W-1} (q_{h,k+1} - q_{hk})^2 + \frac{1}{(H-1)W} \sum_{h=1}^{H-1} \sum_{k=1}^{W} (q_{h+1,k} - q_{hk})^2, \tag{28}$$

$$V(q) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{k=1}^{W} (q_{hk} - \bar{q})^2, \qquad G(q) = \frac{E(q)}{V(q)}.$$

Horizontal and vertical differences are averaged separately over their valid adjacent-pixel pairs, without boundary padding. The ratio is undefined for $V(q) \le 10^{-12}$; all displayed measurements have variance above this threshold.

All retained trajectories. We summarize all 20 directed trajectories from the ten selected image pairs, using all nine interpolation values per trajectory. For a statistic $m \in \{\overline{Y}, G\}$, let $m_\alpha = m(D_R(P_r, R_\alpha))$ and define

$$g_m = m(D_R(P_d, R_d)) - m_0, \qquad \Delta_m = m_1 - m_0.$$

Thus, donor direction is determined by the donor's versus receiver's native reconstruction, not their source photographs. For $|g_m| > 10^{-12}$, set $s_m = \operatorname{sign}(g_m)$. An endpoint is aligned when $s_m \Delta_m > 10^{-12}$. A donor-directed monotonic trajectory satisfies

$$s_m(m_{(j+1)/8} - m_{j/8}) \ge -10^{-12} \quad \text{for all } j = 0, \ldots, 7.$$

Luminance follows the donor-defined direction at 19/20 endpoints, with 14/20 donor-directed monotonic trajectories. Normalized gradient energy follows that direction at 20/20 endpoints, with 17/20 monotonic trajectories. The opposite luminance response, `p08_ab`, remains included. One gradient-energy response overshoots the donor statistic. No neutral endpoints or degenerate donor-direction differences occur.

These counts describe the retained image pairs. Interpolation steps, reciprocal directions, and pairs sharing source images are not independent experimental replicates; no population-level confidence intervals are assigned. Directional alignment does not require reaching the donor statistic or preserving every other image property.

G.5 WAVELET DISTANCE

FWD uses the authors' Haar level-four wavelet-packet implementation (Veeramacheneni et al., 2025). For the three-seed ImageNet-100 appearance experiment, mean FWD is 0.86126 for patch-only decoding and 0.83224 for register-input decoding. The paired change is -0.02902, with an unadjusted 95% seed-t interval of $[-0.07072, 0.01267]$. The register-input model has lower FWD in each seed, but the interval includes zero.

H Additional Reconstruction and Generation Examples

Figures 6 and 7 show six reconstruction examples and four generated-image pairs, respectively. These selections use distinct classes within each gallery and exclude image identities shown in Figure 4. All displayed images use ImageNet-1K-trained DINOv2 ViT-B decoders; model and sample identities are given in Appendix F.6.

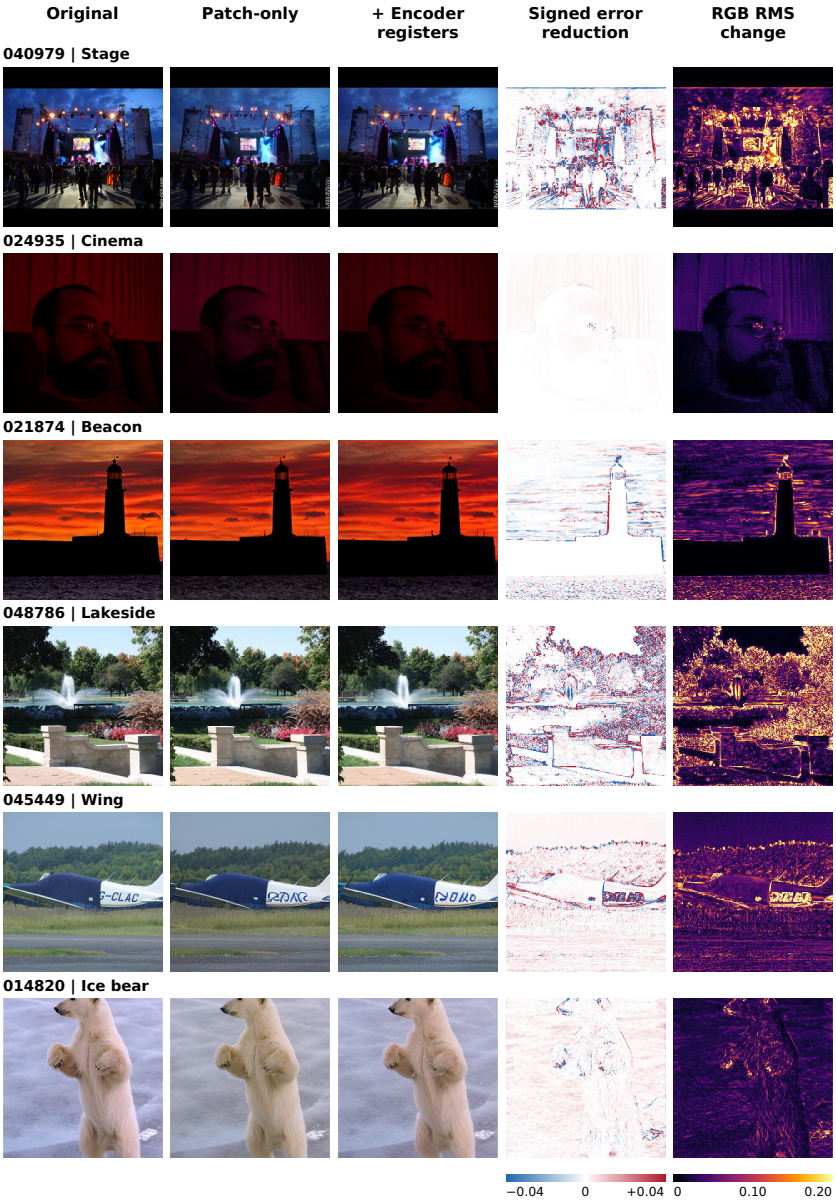

Figure 6: **Additional ImageNet-1K reconstructions.** Columns show the original image, patch-only reconstruction, reconstruction with four encoder registers, signed RGB MSE reduction, and RGB RMS change. Signed maps show patch-only minus register-input squared error against the original, averaged over RGB channels: red indicates lower error with registers and blue lower error with patches alone. RMS maps show unsigned differences between reconstructions. Fixed scales are $[-0.04, 0.04]$ for signed error reduction and $[0, 0.20]$ for RMS change; values outside these ranges saturate for display.

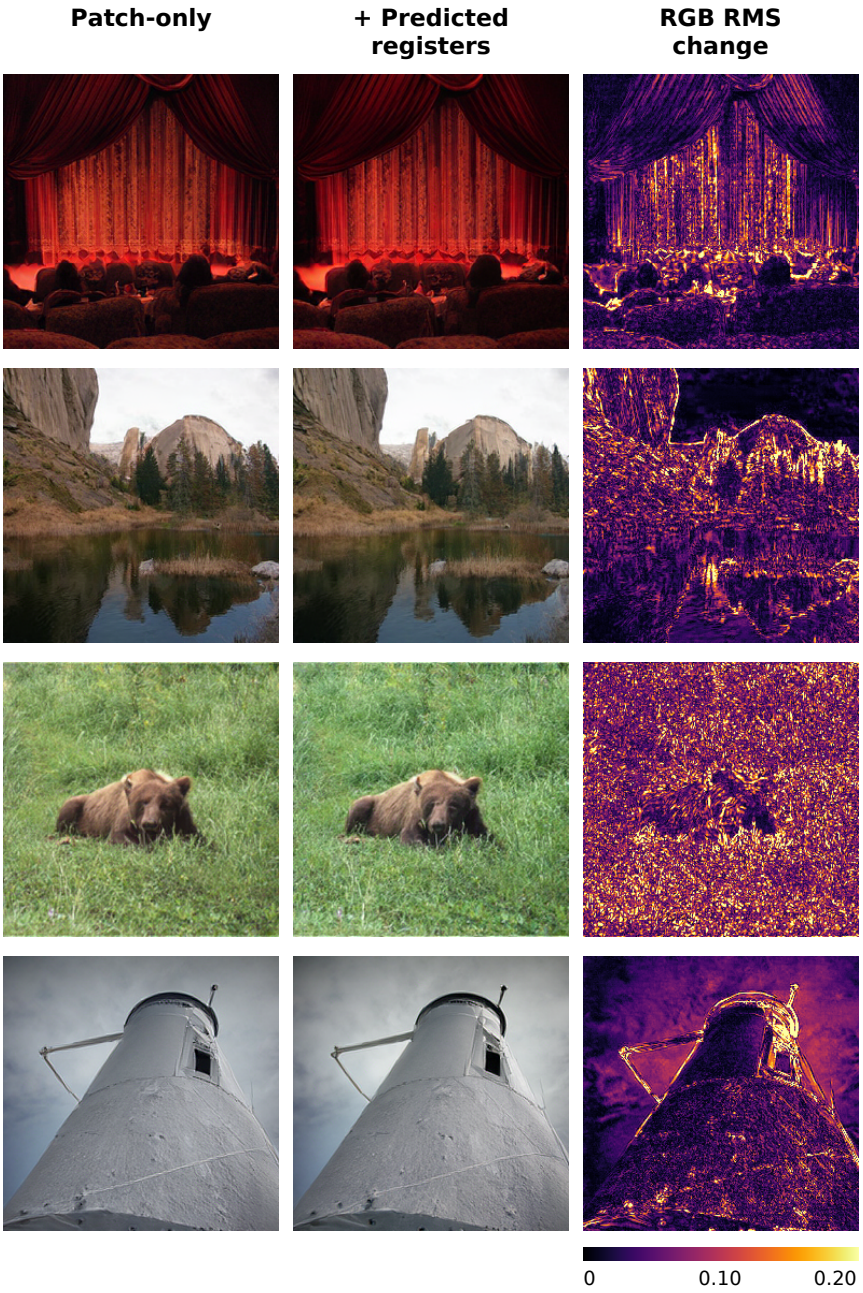

Figure 7: **Additional generation examples.** Each row decodes the same stored pretrained DiT-DH/XL patch tensor using a patch-only decoder or a decoder supplied with MHAP-predicted registers. The right column shows the per-pixel RGB RMS difference between outputs on a fixed $[0, 0.20]$ scale. Brighter values indicate larger output changes; values above 0.20 saturate for display. These maps measure change magnitude, not a difference in generation quality.