SEMANTIC FIELDS ARE INTERFACES: CONSUMER-TYPED FUNCTIONAL MIGRATION FOR REPRESENTATION AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation autoencoders (RAEs) use pretrained visual features as generative states. These dense fields also support retrieval and downstream heads, forming shared interfaces whose consumers may persist across producer releases. Producer upgrades may preserve tensor shape while changing coordinate semantics, rendering retained consumers incompatible. We study post-release field migration, in which an already-trained successor serves retained consumers through one release-wide map while the generator, representation models, and consumers remain frozen. Consumer-Typed Functional Migration (CTFM) learns corrections around a host-specific calibrated base and supervises repair with archived consumer responses rather than field matching alone. Its contract separates behavior to recover from coordinates to preserve: IQ-SAFE repairs decoding while exactly preserving the base-calibrated pooled coordinate, whereas QUALITY permits full-field correction. RAEv2 optionally employs an ownership-preserving, affine-initialized low-rank refinement. Across DINO-based RAEv2 and Perception Encoder (PE) RAEs, CTFM improves recognition, fidelity, semantic consistency, and distribution quality over matched post-hoc alternatives. On 3,232 new RAEv2 fields, a 1.05M-parameter map outperforms identical-capacity feature matching on all five primary metrics. With one quarter of the refinement samples, CTFM gains 3.87 ResNet-50 points and 1.82 dB PSNR over full-data matching; with half the PE residual-fitting classes, it gains 10.88 ResNet-50 points over the feature-to-task procedure. On 7,200 real-image queries, IQ-SAFE preserves every calibrated-base retrieval decision while improving decoded PSNR by 5.95 dB. These results support consumer-defined compatibility for practical reuse of dense RAE fields under fixed consumer contracts.

1 INTRODUCTION

Latent image generation maps noise and conditioning to a representation, which a matched decoder converts into an image (Rombach et al., 2022; Peebles & Xie, 2023; Lipman et al., 2023). The encoder constructs training representations but is absent during noise-to-image inference. Representation autoencoders (RAEs) use pretrained visual features as the latent: an encoder produces a grid of feature vectors, or *semantic field*; a decoder reconstructs images; and a generator learns the field distribution (Zheng et al., 2026; Singh et al., 2026; Tong et al., 2026; Shi et al., 2025; Leng et al., 2026). Fig. 1(a) shows encoded or generated fields feeding the retained decoder and gallery through a shared interface. Semantic features support retrieval and can feed task heads (Bolya et al., 2025). Related autoregressive, object-centric, and video generators expose reusable representations (Yu et al., 2026; Chapin et al., 2026; Guo et al., 2026). RAEs thus **turn a generative latent into a shared semantic interface**, with consumers that can be deployed and maintained independently.

Producer and consumer releases need not coincide (Shen et al., 2020; Hu et al., 2022b; Ran et al., 2024). A successor may use a different representation and its own matched decoder while earlier consumers remain installed. Decoder replacement can require client updates; gallery replacement can require embedding recomputation (Jaeckle et al., 2023); task-head replacement can require training data or coordination across system owners. We study **a producer upgrade without a simultaneous consumer upgrade**. The incompatibility lies at the interface: fields with identical 256×1024

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

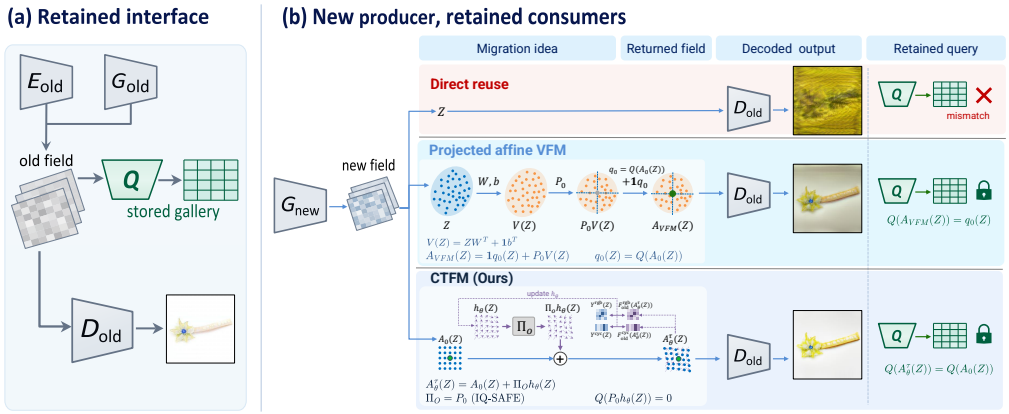


Figure 1: **One field, two obligations: repair decoding, preserve the query.** (a) Retained decoding and retrieval share an archived field interface. (b) Direct reuse, projected affine VFM, and CTFM feed the same D_{old}, using one held-out field and the matched-native image in (a). Projected VFM and IQ-SAFE preserve $q_0(Z) = Q(A_0(Z))$; CTFM repairs RGB and semantic-cycle responses. Fig. 2 details the map summarized by the CTFM row.

dimensions can have different coordinate meanings. Tensor shape alone therefore cannot certify compatibility.

On 648 held-out RAEv2 classes, direct reuse under the archived decoder yields 0.15% gallery-based recognition of decoded images, versus 56.29% under the matched successor decoder. Post-release migration translates the available successor's fields while keeping the producer and installed consumers fixed. The successor decoder supplies offline references; the archived decoder remains the serving renderer. Full-stack replacement and per-consumer routing serve different deployment requirements. In Fig. 1(b), one field must satisfy two requirements: repair decoder-side behavior (F), and preserve the calibrated gallery query through coordinate ownership (O). The decoder can depend on both mean and spatial variation; ownership restricts only the added correction. The projected affine VFM (Mai et al., 2026) and CTFM rows preserve the same calibrated query, separating query preservation from decoded-image quality.

CTFM separates base calibration, compact functional repair, and optional scalable refinement (Fig. 2). A host-specific base A_0 first translates pooled and spatial coordinates. A learned residual then repairs consumer responses around that frozen base. IQ-SAFE removes the residual's token mean, preserving the calibrated query exactly; QUALITY permits full-field motion. Both RAE hosts use compact repair; RAEv2 additionally refines two low-rank factors around the frozen compact map.

Contributions. (i) We formulate post-release dense-field compatibility through consumer behavior and coordinate ownership. (ii) We develop consumer-response supervision with exact pooled-coordinate preservation and an ownership-preserving low-rank extension. (iii) Two independent RAE release pairs establish gains over matched-capacity controls in recognition, fidelity, semantic consistency, and distribution quality. Reduced-data fitting, zero-refit transfer, regional ownership, and response ablations establish efficiency and explain the gains.

2 POST-RELEASE MIGRATION WITH RETAINED CONSUMERS

2.1 AVAILABLE MODELS AND RETAINED CONSUMERS

The successor is available before migration fitting. Encoder E_{new} defines the field representation, decoder D_{new} reconstructs images, and frozen generator G_{new} produces $Z = G_{\text{new}}(\epsilon, c) \sim \mathcal{P}_{\text{new}}$, with noise ϵ, conditioning c, and output distribution $\mathcal{P}_{\text{new}}$. Our principal experiments use class conditioning and fields $\mathcal{H} = \mathbb{R}^{n \times d}$ with n spatial tokens and d channels. The retained consumers include D_{old} and optionally $S_{\text{old}} = \psi \circ Q$, where Q averages tokens and ψ performs gallery scoring or another downstream computation. E_{old} is available offline. The producer and representation

models form a host system. A *release graph* specifies the old/new pairing and available computations. Evaluated pairs use released components with equal field dimensions; generators are not retrained.

A map $A : \mathcal{H} \to \mathcal{H}_{\text{old}}$ must return a field executable by the retained system. For each calibration field, the matched-native image (or realization) is $x_\star(Z) = D_{\text{new}}(Z)$: the reference is obtained by decoding the same field being migrated. The image or a fixed archived computation defines $Y(Z)$; computations requiring archived fields receive $E_{\text{old}}(x_\star(Z))$. Thus offline pairs $(Z, Y(Z))$ need no manual cross-release correspondence, and image recovery requires $D_{\text{old}}(A(Z)) \approx D_{\text{new}}(Z)$.

All generator, encoder, and decoder weights remain fixed during migration; change across releases precedes learning during migration. Frozen models transmit gradients to the map without weight updates. Deployment evaluates A followed by installed consumers, with no additional native decoding, archived re-encoding, candidate search, or per-input optimization. Real-image transfer applies the same map to $Z = E_{\text{new}}(x)$, testing generalization to shifted real-image inputs. Each generated field is one state, shared across the compared maps.

2.2 Consumer-typed interface contracts

A retained gallery reads the token average, while decoder repair can change spatial variation. Mean pooling and token centering separate the shared global coordinate from the spatial component:

$$Q(Z) = \frac{1}{n}\mathbf{1}^\top Z, \qquad P_0 Z = Z - \mathbf{1}Q(Z). \tag{1}$$

Here $\mathbf{1} \in \mathbb{R}^n$ contains ones. The decomposition $Z = \mathbf{1}Q(Z) + P_0 Z$ separates the pooled coordinate and mean-zero spatial component. Corrections in $\ker Q$ preserve mean-pooled inputs while allowing decoder responses to change through spatial and channelwise adjustments.

A consumer-typed interface contract $\tau = (F, O)$ pairs the archived computation F to recover with a linear statistic O to preserve. These roles can belong to different consumers: decoder-side responses define F, while $O = Q$ protects the gallery input. The release chooses both according to its installed consumers; $O = 0$ permits unrestricted correction. First fit a base translation A_0. The residual form $A(Z) = A_0(Z) + H(Z)$ separates calibration from further repair: linearity gives $O(A(Z)) = O(A_0(Z))$ whenever $O(H(Z)) = 0$. Calibration sets the protected pooled coordinate $q_0(Z) = Q(A_0(Z))$; with $O = Q$, residual repair preserves $q_0(Z)$ while improving the declared response. The gallery reads the returned field directly, independently of image decoding.

Definition 1 (post-release consumer-typed migration). Given a fixed base A_0, frozen producer and consumers, and offline paired responses, learn one release-wide map $A(Z) = A_0(Z) + H(Z)$ minimizing functional mismatch subject to $H(Z) \in \ker O$ for every Z. At deployment, the map makes no additional archived-model calls; retained consumers execute $A(Z)$.

2.3 Relation to prior compatibility work

Representation compatibility. Backward-compatible embeddings preserve vector comparability or backfill stored features (Shen et al., 2020; Zhou et al., 2023; Zhang et al., 2022; Hu et al., 2022b; Ricci et al., 2024; 2025; Ramanujan et al., 2022; Jaeckle et al., 2023; Yoon & Arik, 2025; Vejendla, 2025). Extensions address cross-modal retrieval, online backfilling, uncertainty, and frozen-query adaptation (Jang & Lim, 2025; Seo et al., 2025; Bui et al., 2025; Zhao, 2026). Our post-release setting additionally requires a dense field executable by an immutable image decoder.

Stitching and generative adapters. Stitching connects neural representations, including vision foundation models (Badrinath et al., 2024; Traft, 2026; Balogh & Jelasity, 2025; Mai et al., 2026; Liu et al., 2026b; Wu & Zhu, 2026). CTFM combines functional alignment and invariance (Smith et al., 2025; Athanasiadis et al., 2026) at a frozen dense-field interface. Generative bridges, decoder branches, and inversion change the receiving graph or optimize per input (Ran et al., 2024; Mao et al., 2026; Jian et al., 2026; Allakhverdov et al., 2026; Ouyang & Mao, 2025; Li et al., 2025). Tokenizer/RAE redesigns change the representation stack (Chen et al., 2026; Liu et al., 2026a; Zhu et al., 2026; Huang et al., 2026; Zheng et al., 2025); dense-feature adaptations and single-consumer stitches address other interfaces (Wandel et al., 2026; Wang et al., 2026; Zhang et al., 2026; Chaudhury, 2026; Li et al., 2026). CTFM distills responses (Hinton et al., 2015; Romero et al., 2015) into an ownership-constrained field map with frozen producer and consumers (App. C).

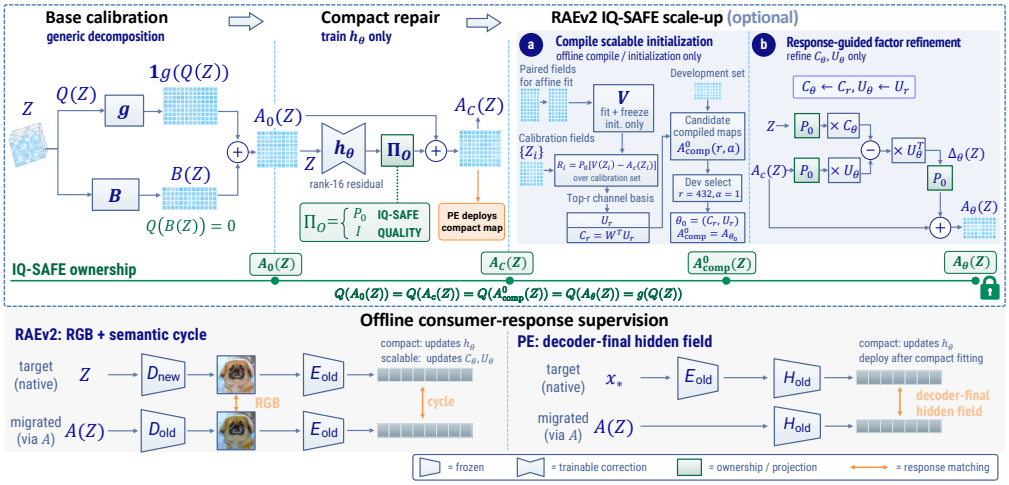


Figure 2: **CTFM: calibrate, repair, and optionally refine.** The upper pipeline calibrates A_0 and fits a host-specific compact residual; PE then deploys. RAEv2 optionally (a) selects a compiled initialization and (b) refines only C_θ, U_θ around frozen A_c. The green rail tracks the protected pooled coordinate. Below, RAEv2 uses RGB then cycle supervision for compact fitting, and both for refinement; PE uses decoder-final features. Frozen consumers transmit gradients to the fitted residual or factors.

3 Consumer-Typed Functional Migration

3.1 Base calibration and compact functional repair

Fig. 2 separates map construction (top) from offline consumer-response supervision (bottom). Base calibration fits A_0 to paired successor fields and archived encodings of matched-native images, providing a starting point for residual repair. The pooled map g translates token means into archived coordinates; $B(Z)$ supplies the mean-zero spatial component:

$$A_0(Z) \triangleq \mathbf{1}g(Q(Z)) + B(Z), \qquad Q(B(Z)) = 0. \tag{2}$$

CTFM permits host-specific base estimators: RAEv2 uses ridge transport with a rank-32 correction conditioned on the pooled input; PE uses pooled and centered transports with a pre-fitted residual. Base parameters and standardization statistics are frozen before residual fitting and shared by matched controls (App. B). The same base sets the protected pooled coordinate in each matched comparison.

Archived response targets. After an RGB warm-start, RAEv2 compact fitting matches the semantic cycle $F^{\mathrm{cyc}}_{\mathrm{old}}(X) = E_{\mathrm{old}}(D_{\mathrm{old}}(X))$ to $E_{\mathrm{old}}(x_\star(Z))$. Scalable refinement additionally matches $F^{\mathrm{rgb}}_{\mathrm{old}}(X) = D_{\mathrm{old}}(X)$ to $x_\star(Z)$. The second RAE host uses vision-language Perception Encoder (PE) features (Bolya et al., 2025). Its functional is the archived decoder's final hidden field $H_{\mathrm{old}}(X)$, with target $H_{\mathrm{old}}(E_{\mathrm{old}}(x_\star(Z)))$ (App. B).

Trainable compact correction. Only h_θ is optimized, using token exchange and pooled-feature conditioning. RAEv2 uses one rank-32 branch with a zero-initialized output; PE uses two rank-16 branches with cancelling outputs. Both start at the calibrated base with 163,840 weights (Apps. A and H.8). PE deploys this compact map; RAEv2 adds optional refinement (Sec. 3.3). Let Π_O project onto $\ker O$ and $\mathcal{P}_{\mathrm{cal}}$ denote the calibration distribution. For K responses $F_{\mathrm{old}} = (F^1_{\mathrm{old}}, \dots, F^K_{\mathrm{old}})$ and targets $Y = (Y^1, \dots, Y^K)$, vectorize tensor responses for the norm. With displacement weight $\lambda \geq 0$, the map and its offline fitting objective are

$$A^\tau_\theta(Z) = A_0(Z) + \Pi_O h_\theta(Z), \tag{3}$$

$$\theta^* = \arg\min_\theta \mathbb{E}_{Z\sim\mathcal{P}_{\mathrm{cal}}} \left[\|F_{\mathrm{old}}(A^\tau_\theta(Z)) - Y(Z)\|_2^2 + \lambda \|A^\tau_\theta(Z) - A_0(Z)\|_F^2 \right]. \tag{4}$$

Both A_0 and h_θ receive Z; only the correction is projected before summation. The objective balances archived-response mismatch against displacement from the calibrated base. Host-specific compact loss normalizations are given in App. B; Sec. 3.3 defines scalable refinement.

3.2 EXACT COORDINATE OWNERSHIP

The two contracts differ in permissible correction directions: IQ-SAFE uses $O = Q$ and $\Pi_O = P_0$; QUALITY uses $O = 0$ and $\Pi_O = I$.

Proposition 1 (exact owned-coordinate invariance). For every field Z and residual h_θ,

$$Q(A_\theta^{\text{IQ-SAFE}}(Z)) = g(Q(Z)). \tag{5}$$

Linearity proves the identity (App. A), protecting every deterministic $\psi \circ Q$, including pooled linear or MLP classifiers. Response-based heads inherit error bounds under the coverage conditions (Apps. A.3 and H). Finite-precision retrieval is checked using a sufficient condition comparing score perturbations with the top-two margin (App. H.11). Any linear ownership operator with a computable nullspace projector admits the same construction; Sec. 4.3 tests regional means.

Ownership geometry. For a desired correction R, constant-token and mean-zero components are orthogonal:

$$\|R\|_F^2 = \|P_0 R\|_F^2 + n\|Q(R)\|_2^2. \tag{6}$$

Thus $P_0 R$ is the nearest IQ-SAFE correction, and the last term measures excluded energy. Because a consumer can respond differently to the two components, exact ownership can affect attainable response recovery. App. A derives the corresponding first-order response bound; Sec. 4.3 evaluates the ownership intervention, varying ownership with the fitted map held fixed.

3.3 OWNERSHIP-PRESERVING LOW-RANK EXPANSION

After calibrating A_0 and fitting the compact residual, RAEv2 freezes the resulting map A_c as an anchor. The optional scalable stage constructs A^0_{comp} and refines only two factors to obtain A_θ. Each map receives Z. The stage arrows in Fig. 2 denote offline construction; deployment combines the anchor and projected correction for the same incoming field. Panel (a) constructs the shared initialization; panel (b) updates its two factors through frozen consumer responses.

Let V be a frozen affine aligner fitted offline to paired fields (Affine VFM). The displacement $V(Z) - A_c(Z)$ is centered and restricted to r channel directions: the orthonormal columns of $U_r \in \mathbb{R}^{d\times r}$, estimated from centered calibration displacements. The ownership-feasible tangent $\mathcal{T}_r = \{XU_r^\top : X \in \mathbb{R}^{n\times r}, Q(X) = 0\}$ has projector $\Pi_{\mathcal{T}_r}(R) = P_0 R U_r U_r^\top$: left projection preserves token ownership; right projection limits channel capacity. With displacement scale α,

$$A^0_{\text{comp}}(Z) = A_c(Z) + \alpha \Pi_{\mathcal{T}_r}(V(Z) - A_c(Z))\,. \tag{7}$$

For $V(Z) = ZW^\top + \mathbf{1}b^\top$, centering cancels the bias and $C_r = W^\top U_r$ gives an implementation with two $d \times r$ factors, without a per-input coefficient predictor:

$$\begin{aligned} \Delta^0_r(Z) &= \alpha\Big[(P_0 Z)C_r - (P_0 A_c(Z))U_r\Big]U_r^\top, \\ A^0_{\text{comp}}(Z) &= A_c(Z) + P_0\Delta^0_r(Z). \end{aligned} \tag{8}$$

Proposition 2 (exact ownership preservation under compilation). For every Z, $Q(A^0_{\text{comp}}(Z)) = Q(A_c(Z)) = g(Q(Z))$. For affine V, the correction is $\alpha\Pi_{\mathcal{T}_r}(V(Z) - A_c(Z))$.

Centering and projection establish Proposition 2 (App. A.5). A fixed labeled development grid selects $\alpha = 1$ and $r = 432$, the largest candidate within the affine weight budget. The map's 1,048,576 weights comprise the frozen 163,840-weight anchor and 884,736 factor weights, 1,024 fewer than full affine VFM. Both factors are stored once and shared across all deployment inputs.

Response-guided factor refinement. Initialize $\theta = (C_\theta, U_\theta)$ at (C_r, U_r) and replace the factors in Eq. (8) to obtain Δ_θ and $A_\theta(Z) = A_c(Z) + P_0\Delta_\theta(Z)$. The same U_θ is used in subtraction and expansion. The anchor, rank, and projector remain fixed.

For each response $k \in \{\text{rgb}, \text{cyc}\}$ defined in Sec. 3.1, let $e_k(Z;\theta)$ denote the response MSE. Fix $s_k(Z) = e_k(Z;\theta_0)$ at compiled initialization and minimize the larger normalized error for each

sample. For minibatch $\mathcal{B}$, define $\mathrm{mse}_{\mathcal{B}}[X] = (|\mathcal{B}|nd)^{-1} \sum_{Z\in\mathcal{B}} \|X(Z)\|_F^2$. The implemented loss is

$$\mathcal{L}^{\mathcal{B}}_{\mathrm{resp}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{Z\in\mathcal{B}} \max_{k\in\{\mathrm{rgb,cyc}\}} \frac{e_k(Z;\theta)}{\max\{s_k(Z), 10^{-8}\}} + \lambda \frac{\mathrm{mse}_{\mathcal{B}}[A_\theta - A^0_{\mathrm{comp}}]}{\max\{\mathrm{mse}_{\mathcal{B}}[A^0_{\mathrm{comp}}], 10^{-8}\}}. \tag{9}$$

The denominators remain fixed. The maximum prioritizes the response with the larger fraction of its initial error remaining. The compact regularizer uses A_0 as its reference; factor refinement measures motion from A^0_{comp}. Centering preserves the ownership identity in Eq. (8) throughout refinement; the orthogonal channel projection interpretation applies only at initialization. The principal control refines the same factors with feature matching under identical optimization conditions. App. D.6 isolates response and normalization choices; App. A.5 gives implementation details.

3.4 Fitting and deployment procedure

Consumers, ownership, and stage-specific splits define offline calibration and repair. Labeled development selects initialization; response risk selects the refinement checkpoint before evaluation. App. B distinguishes sample and class independence. Deployment evaluates the map and installed consumers (Alg. 1). PE's 163,840 residual weights occupy a 0.640-MiB serialized checkpoint. Rank-432 RAEv2 uses 4.00 MiB of raw weights and 9.06 MB including serialized dependencies (binary MiB; decimal MB). At batch eight, the map adds 0.079 ms per sample, or 1.04% of decoder-only latency. App. G separates learned state, complete dependencies, operations, and end-to-end cost.

4 Experiments

We evaluate archived decoding and protected retrieval on frozen successor fields, then test the same maps on real-image inputs. System comparisons measure compatibility; matched refinements isolate supervision; composition and transfer test reuse beyond the fitting setting.

4.1 Experimental protocol

Hosts and samples. RAEv2 pairs a DINOv3-L generator/decoder with a DINOv2-L reader (Singh et al., 2026; Siméoni et al., 2025; Oquab et al., 2023). The second RAE host tests migration beyond DINO representations. PE-Core targets global recognition; PE-Spatial aligns features for dense prediction (Bolya et al., 2025). We pair DiffusionBench's PE-Spatial-L generator/decoder with a PE-Core-L reader (Leng et al., 2026). Both hosts expose 256×1024 fields; only the pairing and map are constructed here. Tab. 1 specifies evaluation sizes. System sets contain 2,592 RAEv2 and 3,200 PE generated samples; separate prospective sets contain 3,232 and 1,600. The RAEv2 prospective set uses fresh seeds but includes 160 earlier selection classes; excluding the 32 compiler-selection classes retains all seven gains on 776 classes (Tab. 35). System and PE evaluation classes exclude their fitting/selection sets (App. B). All methods share evaluation states and references; Tab. 6 specifies the stage-specific splits.

Functionals and matched controls. RAEv2 isolates RGB-plus-cycle versus feature supervision. PE tests a procedure: decoder-final feature matching (FFM) versus FFM followed by gallery task loss (FFM-to-TLT). Functionals are fixed before evaluation (App. B). Principal RAEv2 refinements share exactly 1,048,576 map weights, initialization, data order, optimizer, projector, and update budget. Full affine VFM has 1,049,600 weights; compact comparisons use 163,840. LoRA controls update decoder weights (Hu et al., 2022a). Shared scalable initialization uses labeled development data, but refinement and checkpoint selection do not use class labels.

Table 1: **Two RAE release pairs and evaluation populations.** Arrows map successor to archived coordinates. System sets compare maps; prospective sets test objectives and fitting breadth. App. B specifies class independence.

Host	Release pair	Fitted response	Evaluation samples
RAEv2	DINOv3-L → DINOv2-L	RGB + semantic cycle	system: 648×4; prospective: 808×4
PE	Spatial-L → Core-L	decoder-final field	system: 800×4; prospective: 800×2

Table 2: **RAEv2 ownership-constrained frontier**: 2,592 states from 648 held-out classes. Weights exclude the shared base. Drift is maximum pooled change from that base in normalized archived coordinates. Scalable refinements share anchor, rank, and capacity, differing only in objective. Identical-image PSNR is capped at 120 dB; bold marks the best comparable map.

Method	Map weights	Gallery ↑	ResNet-50 ↑	PSNR ↑	LPIPS ↓	Cycle ↓	Drift ↓
Without an additional learned residual							
Direct reuse	0	0.15	0.19	9.398	0.8293	1.8515	2.99
Calibrated base A_0	0	10.03	7.99	11.595	0.6384	1.5437	**0.00**
Compact IQ-SAFE field maps							
Feature matching (VFM)	163.8K	21.99	20.72	13.802	**0.5236**	1.3556	$< 10^{-6}$
Feature matching + task loss	163.8K	22.65	21.33	13.847	0.5241	1.3457	$< 10^{-6}$
Task-only state map	163.8K	11.03	8.53	11.628	0.6344	1.5295	$< 10^{-6}$
CTFM anchor	163.8K	**35.76**	**37.27**	**16.505**	0.5237	**1.0995**	$< 10^{-6}$
Graph-changing controls							
Semantic decoder LoRA	163.2K	12.15	1.85	11.201	0.8466	1.4262	–
Pixel decoder LoRA	163.2K	9.99	2.12	12.782	0.6945	1.5843	–
Relaxed decoder head	885.5K	13.04	1.58	10.966	0.9254	1.4150	–
Scalable IQ-SAFE field maps							
Projected affine VFM	1,049.6K	52.70	75.39	17.256	0.2408	0.5641	$< 10^{-6}$
Compiler initialization	1,048.6K	52.66	74.88	16.967	0.2546	0.6009	$< 10^{-6}$
Feature-matching refinement	1,048.6K	52.04	71.84	15.939	0.2927	0.6568	$< 10^{-6}$
CTFM response refinement	1,048.6K	**53.28**	**75.96**	**17.816**	**0.2376**	**0.5315**	$< 10^{-6}$
Ownership-free affine reference							
Affine VFM, QUALITY	1,049.6K	53.90	78.55	18.670	0.1969	0.5029	5.65
Reference conditions							
Archived field	0	55.71	80.86	19.482	0.1438	0.2985	–
Matched-native image	0	56.29	81.02	120.000	0.0000	0.0030	–

We reproduce the stitching objectives of Mai et al. (2026) (VFM baselines) in 144 fits across two hosts, three map families, four objectives, three learning rates, and two weight decays (Tab. 9). Without executable training code at the 30 August 2026 cutoff, we reproduce the objectives independently. App. C distinguishes alternatives by training, mutable components, and deployment access.

Endpoints. Decoded-gallery Top-1 measures target-class accuracy by classifying mean-pooled $E_{\text{old}}(D_{\text{old}}(A(Z)))$ against the stored gallery. This accuracy can improve while direct pooled retrieval using $Q(A(Z))$ remains unchanged. ResNet-50 (He et al., 2016) evaluates decoded images; independent evaluators EfficientNetV2-S and ViT-B/16 (Tan & Le, 2021; Dosovitskiy et al., 2021) are excluded from fitting and selection. PSNR measures pixel fidelity, LPIPS perceptual distance (Zhang et al., 2018), and cycle MSE archived re-encoding error. Image/feature references are $(D_{\text{new}}(Z), E_{\text{old}}(D_{\text{new}}(Z)))$ for generated fields and $(x, E_{\text{old}}(x))$ for real images, with $Z = E_{\text{new}}(x)$ and identical preprocessing. Archived-field references use exact E_{old} outputs; matched-native images use the successor decoder. Cache precision can produce a nonzero native cycle floor (App. B). Tab. 4(a) reports direct retrieval (R@1/R@5/mAP) alongside decoding; (b) reports native- and class-matched real-reference FID (Heusel et al., 2017), paired Inception-feature MSE, and classwise symmetric nearest-neighbor distance (adequacy). App. D.5 defines these distribution metrics and KID (Bińkowski et al., 2018); Apps. B to D provide full values and uncertainty. Blue/orange/gray rows denote CTFM/ownership-free alignment/references.

4.2 COMPATIBILITY UNDER THE DECLARED CONSTRAINT

System performance and objective control. Compact CTFM outperforms matched-capacity maps on Gallery, ResNet-50, PSNR, and cycle, with LPIPS within 0.0001 of the best (Tab. 2). Rank-432 CTFM improves all five endpoints over projected affine VFM. Decoder-changing controls yield lower semantic accuracy under the evaluated adaptation budget. Relative to direct reuse, final response refinement closes 95.6% of the gap to the archived-field reference in decoded-gallery Top-1, 93.9% in ResNet-50 Top-1, and 83.5–86.3% across PSNR, LPIPS, and cycle error. Under identical initialization and optimization, response refinement improves recognition and fidelity, whereas feature matching reduces both (optimizer-order repeats in Tab. 24). Excluding the 32 compiler-selection classes leaves 776 classes on which CTFM retains gains of 3.64 ResNet-50 points and 1.86 dB

Table 3: **Consumer-functional supervision with fixed or reduced data.** Equal-capacity blocks share frozen transport and use matched, 25% RAEv2, or 50% PE residual-stage data. RAEv2 blocks use different evaluation seeds. Independent is mean EfficientNetV2-S/ViT-B/16 Top-1.

Method	Fit classes	Gallery ↑	ResNet-50 ↑	PSNR ↑	LPIPS ↓	Cycle ↓	Indep. ↑
RAEv2: matched refinement data							
Feature matching	160	52.970	72.246	15.885	0.29071	0.65059	74.489
CTFM	160	**54.548**	**75.650**	**17.744**	**0.23830**	**0.52842**	**77.491**
RAEv2: CTFM uses 25% refinement data							
Feature matching	160	55.229	73.082	15.876	0.29313	0.65389	75.866
CTFM	40	**56.869**	**76.949**	**17.699**	**0.24725**	**0.54969**	**78.605**
PE: CTFM uses 50% residual-fit data							
Feature matching + task loss	128	25.875	48.313	17.059	0.49537	18.74895	53.156
CTFM	64	**27.937**	**59.188**	**17.184**	**0.40479**	**18.43505**	**63.000**

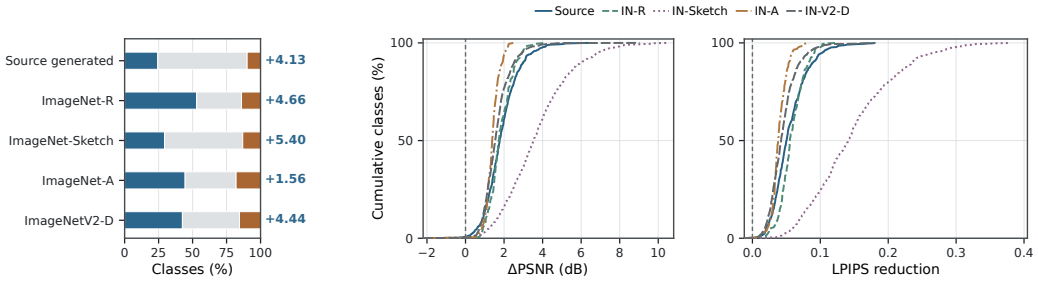
Table 4: **Operational preservation and cross-host distribution quality.** (a) Mean of the two domain means on 7,200 ImageNet-R/Sketch queries; bold marks best ownership-preserving values and ties. (b) Percentages denote residual-stage data; paired MSE and adequacy are Inception-feature distances. RAEv2 uses rank 432, PE the compact map. Tabs. 15 and 17 give complete results.

(a) Executed consumer

Field	R@1↑	R@5↑	mAP↑	PSNR↑	LPIPS↓
Direct reuse	0.11	0.46	0.72	8.237	0.8084
Calibrated base	**5.22**	**15.20**	**11.05**	8.573	0.6944
Proj. affine VFM	**5.22**	**15.20**	**11.05**	12.981	0.3185
IQ-SAFE CTFM	**5.22**	**15.20**	**11.05**	**14.522**	**0.2873**
Archived field	30.15	54.94	41.78	17.794	0.1564

(b) Distribution quality

Host	Method/data	Nat. FID↓	Pair MSE↓	Real FID↓	Adeq.↓
RAEv2	Feature match./full	20.2395	0.050885	20.0387	0.092381
	CTFM / 25%	**15.4744**	**0.040945**	**19.3026**	**0.088616**
PE	FFM-to-TLT/full	89.2179	0.123169	65.2921	0.136443
	CTFM / 50%	**74.9794**	**0.105432**	**55.1153**	**0.126148**



(a) Recognition outcomes. (b) PSNR distributions. (c) LPIPS distributions.

Figure 3: **Complete class-level transfer summaries.** (a) Blue/gray/orange denote improved/tied/degraded ResNet-50 class accuracies; right-margin values are mean gains (pp). (b,c) Class-level CDF traces retain every class and tail; positive effects favor CTFM and the dark dashed line marks zero. Checkpoints are shared and capacity-matched. Source uses generated fields; the other populations use real images. V2-D denotes gallery-disjoint ImageNetV2 (App. E).

PSNR over identical-capacity feature matching, without refitting (Tab. 35). At 800 matched updates, joint per-state normalization improves both RGB and cycle on 90.59% of evaluation states, versus 79.21% for population normalization and 51.49% for cycle-only fitting (App. D.6). Fig. 4 illustrates simultaneous semantic and visual repair.

Data efficiency of residual fitting. CTFM improves every reported endpoint over full-data controls with one-quarter RAEv2 refinement data or half PE residual-fitting data (Tab. 3). Only factor-refinement data vary for RAEv2; the base, anchor, affine source, and initialization remain shared. PE reduces residual-fitting classes with a shared base. Fig. 9 shows the PE trade-off. CTFM fitted on 64 classes improves recognition and PSNR over the 128-class control; fitting on 32 classes yields higher recognition at lower PSNR. The PE comparison uses compact repair without low-rank expansion, showing that the data-efficiency advantage also holds outside the scalable RAEv2 pipeline. Improvements in native- and real-reference FID further connect residual-stage data efficiency to distribution quality (Tab. 4).

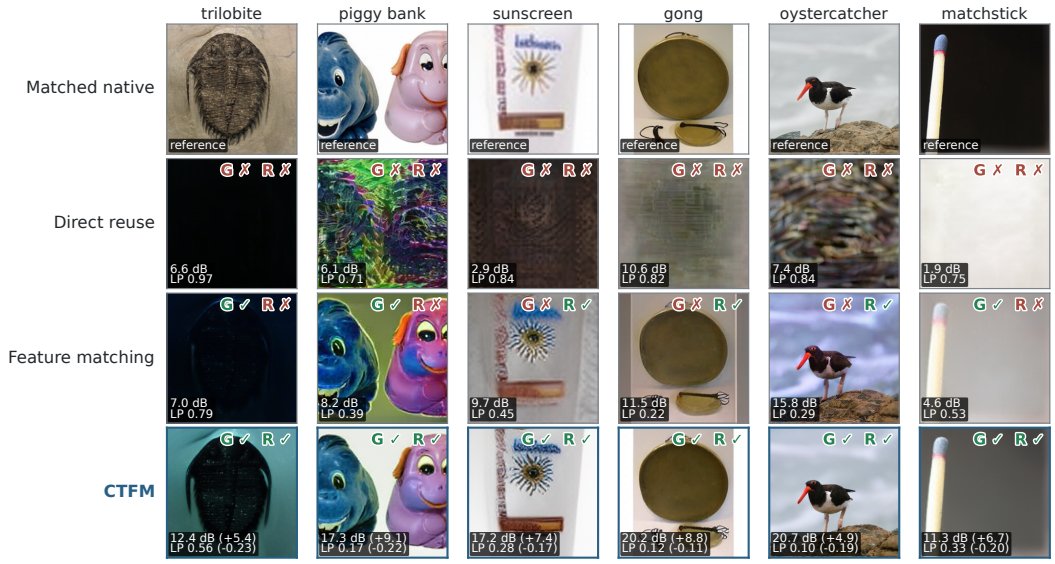


Figure 4: **Six examples of joint semantic and fidelity recovery at identical map capacity.** Direct reuse decodes the successor field without migration. Feature matching and CTFM each contain 1,048,576 map weights. On every displayed state, CTFM is Gallery- and ResNet-50-correct, feature matching misses at least one decision, PSNR rises by at least 4 dB, and LPIPS falls by at least 0.10. G/R denotes Gallery/ResNet-50; ✓/× indicates a correct/incorrect target-class prediction. App. F gives the deterministic selection rule, verified renders, and paired error maps.

4.3 Ownership and Contract Composition

Centering equal regions in a 2×2 or 4×4 partition of the 16×16 grid preserves four or sixteen anchor means without refitting or added weights. On 808 fixed-order samples (one per class), CTFM outperforms same-projector feature matching on all seven endpoints. With sixteen regions, CTFM gains 5.94 ResNet-50 points and 1.41 dB PSNR at maximum drift 1.03×10^{-6}; the four-region map retains 99.69% of the global map's ResNet gain above the anchor (Tab. 13; PE intervention in App. H.10).

Relaxing pooled ownership of a fixed affine map improves fidelity but can change retrieval decisions (Fig. 10). CTFM preserves each base-calibrated query's top-ranked item while improving decoding (Tab. 4). All 7,200 queries satisfy the sufficient score-margin condition, with identical float32/float64 decisions (Tab. 34). The gallery embeddings and scoring rule remain fixed throughout.

4.4 Release-Level Transfer and Deployment

Across generation and four ImageNet shifts, CTFM outperforms matched feature refinement in all 35 mean comparisons (seven metrics per dataset; Fig. 3). Equal-weight dataset averages also favor CTFM over projected affine VFM in Gallery, ResNet-50, independent recognition, PSNR, LPIPS, and cycle. The independently fitted PE map transfers without refitting (App. E). On both hosts, CTFM improves native and real-reference FID and all reported classwise feature distances (Tab. 4). App. H.6 reports KID sensitivity; App. G accounts for stored dependencies and execution overhead.

5 Conclusion

Representation autoencoders expose dense semantic fields to decoding, retrieval, and downstream tasks. CTFM enables producer evolution with retained consumers by matching responses under coordinate-ownership constraints. IQ-SAFE repairs decoding while preserving the protected pooled coordinate; producer and consumer weights remain frozen. Two RAE release pairs demonstrate matched-capacity gains, reduced residual-stage data requirements, and zero-refit transfer. Regional ownership and joint-response refinement extend the formulation to distinct consumer requirements.

Reproducibility statement

The appendix specifies proofs, release graphs, objectives, optimization, selection scope, uncertainty, and costs. Supplementary materials provide executable field-map modules, CPU tests, per-state records, environment and asset specifications, and numerical checks. End-to-end reproduction additionally requires the separately licensed upstream models and datasets, which are identified without redistribution.

Ethics statement

This work concerns benign model-version compatibility and image-generation evaluation. It introduces no offensive-security capability and no new personal data. Dataset and checkpoint licenses are documented in the supplementary manifest.

AI use statement

Generative AI tools assisted with literature classification, manuscript review, results analysis, and language polishing. The research motivation and core methodological concepts were conceived by human researchers. All experimental data originated from code that was executed and manually reviewed. The technical workflow figures in the paper were created by the human authors. The scientific claims made in the paper are supported by original literature, mathematical arguments, and experimental records. The authors are responsible for verifying AI-assisted content and ensuring the accuracy, clear attribution, and reproducibility of the submitted work.

References

Eduard Allakhverdov, Dmitrii Tarasov, Elizaveta Goncharova, and Andrey Kuznetsov. Feature inversion as a lens on vision encoders. In *WACV*, pp. 3598–3605, 2026.

Ioannis Athanasiadis, Anmar Karmush, and Michael Felsberg. Grounding functional similarity by invariance-aware model stitching. In *ICML*, 2026.

Charumathi Badrinath, Usha Bhalla, Alex Oesterling, Suraj Srinivas, and Himabindu Lakkaraju. All roads lead to rome? exploring representational similarities between latent spaces of generative image models. *arXiv preprint arXiv:2407.13449*, 2024.

András Balogh and Márk Jelasity. How not to stitch representations to measure similarity: Task loss matching versus direct matching. In *AAAI*, 2025.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 2022.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ngoc Bui, Menglin Yang, Runjin Chen, Leonardo Neves, Mingxuan Ju, Zhitao Ying, Neil Shah, and Tong Zhao. Learning along the arrow of time: Hyperbolic geometry for backward-compatible representation learning. In *ICML*, 2025.

Alexandre Chapin, Emmanuel Dellandrea, and Liming Chen. Slot-RAE: Streamlining object-centric learning via direct representation auto-encoders. *arXiv preprint arXiv:2607.11196*, 2026.

Archie Chaudhury. Forgetting is not erasure: Recovering latent knowledge via transport keys. *arXiv preprint arXiv:2606.02860*, 2026.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models. In *ICLR*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Zhihao Duan, Ming Lu, Justin Yang, Jiangpeng He, Zhan Ma, and Fengqing Zhu. Towards backward-compatible continual learning of image compression. In *CVPR*, 2024.

Hyojun Go, Hyungjin Chung, Prune Truong, Goutam Bhat, Li Mi, Zhaochong An, Zixiang Zhao, Dominik Narnhofer, Serge Belongie, Federico Tombari, and Konrad Schindler. Stitched value model for diffusion alignment. *arXiv preprint arXiv:2605.19804*, 2026.

Panagiotis D. Grontas, Antonio Terpin, Efe Balta, Raffaello D'Andrea, and John Lygeros. Pinet: Optimizing hard-constrained neural networks with orthogonal projection layers. In *ICLR*, 2026.

Minghui Guo, Shengqiong Wu, and Hao Fei. V-RAE: Rethinking video latent spaces for generation. *arXiv preprint arXiv:2608.13556*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pp. 15262–15271, 2021b.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022a.

Weihua Hu, Rajas Bansal, Kaidi Cao, Nikhil Rao, Karthik Subbian, and Jure Leskovec. Learning backward compatible embeddings. In *KDD*, 2022b.

Yibin Huang, Jixiang Hong, Zongzhao Li, Yuhan Dai, Zhibin Wang, Chunwei Wang, Jun Song, Chen Wang, Xiaofei Sun, Xiaoxiao Xu, and Conghui Zhu. SPAE: Spectrally guided autoencoder for pretrained visual latents. *arXiv preprint arXiv:2608.01306*, 2026.

Florian Jaeckle, Fartash Faghri, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Fastfill: Efficient compatible model update. In *ICLR*, 2023.

Young Kyun Jang and Ser-Nam Lim. Towards cross-modal backward-compatible representation learning for vision-language models. In *ICCV*, pp. 1783–1792, 2025.

Siyong Jian, Siyuan Li, Luyuan Zhang, Zedong Wang, Xin Jin, Ying Li, Cheng Tan, and Huan Wang. RankE: End-to-end post-training for discrete text-to-image generation with decoder co-evolution. *arXiv preprint arXiv:2605.21195*, 2026.

Xingjian Leng, Jaskirat Singh, Zhanhao Liang, Ethan Smith, Martin Bell, Aninda Saha, Yuhui Yuan, and Liang Zheng. Diffusionbench: On holistic evaluation of diffusion transformers. *arXiv preprint arXiv:2606.24888*, 2026.

Yueyuan Li, Yifei Xiao, Mingyang Jiang, Xiang Zuo, Songan Zhang, and Ming Yang. A stitch in time saves nine: Preserving policy compatibility under perception updates in end-to-end autonomous driving. *arXiv preprint arXiv:2606.21509*, 2026.

Zixiang Li, Haoyu Wang, Wei Wang, Chuangchuang Tan, Yunchao Wei, and Yao Zhao. DCI: Dual-conditional inversion for boosting diffusion-based image editing. In *NeurIPS*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

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*, 2026a.

Xiaoze Liu, Ruowang Zhang, Weichen Yu, Siheng Xiong, Liu He, Feijie Wu, Hoin Jung, Matt Fredrikson, Xiaoqian Wang, and Jing Gao. The vision wormhole: Latent-space communication in heterogeneous multi-agent systems. *arXiv preprint arXiv:2602.15382*, 2026b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Zheda Mai, Ke Zhang, Fu-En Wang, Zixiao Ken Wang, Albert Y. C. Chen, Lu Xia, Min Sun, Wei-Lun Chao, and Cheng-Hao Kuo. Revisiting model stitching in the foundation model era. In *CVPR*, 2026.

Dongxing Mao, Alex Jinpeng Wang, Jiahao Tang, Kevin Qinghong Lin, Linjie Li, Zhengyuan Yang, Lijuan Wang, Min Li, and Jingru Tan. Residual decoder adapter: Id-preserving tokenizer adaption for autoregressive text rendering. In *CVPR*, 2026.

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.

Liangyang Ouyang and Jiafeng Mao. LORE: Latent optimization for precise semantic control in rectified flow-based image editing. *arXiv preprint arXiv:2508.03144*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Vivek Ramanujan, Pavan Kumar Anasosalu Vasu, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Forward compatible training for large-scale embedding retrieval systems. In *CVPR*, 2022.

Lingmin Ran, Xiaodong Cun, Jia-Wei Liu, Rui Zhao, Zijie Song, Xintao Wang, Jussi Keppo, and Mike Zheng Shou. X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In *CVPR*, 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019.

Simone Ricci, Niccolò Biondi, Federico Pernici, and Alberto Del Bimbo. Backward-compatible aligned representations via an orthogonal transformation layer. In *European Conference on Computer Vision Workshops*, pp. 451–464, 2024.

Simone Ricci, Niccolò Biondi, Federico Pernici, Ioannis Patras, and Alberto Del Bimbo. λ-orthogonality regularization for compatible representation learning. In *NeurIPS*, 2025.

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

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Debopam Sanyal, Anantharaman Iyer, Alind Khare, Trisha Jain, Akshay Jajoo, Myungjin Lee, Clayton Kerce, and Alexey Tumanov. KLAS: Using similarity to stitch neural networks for improved accuracy-efficiency tradeoffs. In *ICLR*, 2026.

Seonguk Seo, Mustafa Gokhan Uzunbas, Bohyung Han, Sara Cao, and Ser-Nam Lim. Metric compatible training for online backfilling in large-scale retrieval. In *WACV*, pp. 1537–1545, 2025.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *CVPR*, 2020.

Minglei Shi, Haolin Wang, Borui Zhang, Wenzhao Zheng, Bohan Zeng, Ziyang Yuan, Xiaoshi Wu, Yuanxing Zhang, Huan Yang, Xintao Wang, Pengfei Wan, Kun Gai, Jie Zhou, and Jiwen Lu. SVG-T2I: Scaling up text-to-image latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2512.11749*, 2025.

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.

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

Damian Smith, Harvey Mannering, and Antonia Marcu. Functional alignment can mislead: Examining model stitching. In *ICML*, 2025.

Mingxing Tan and Quoc Le. EfficientNetV2: Smaller models and faster training. In *ICML*, 2021.

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.

Neil Traft. Bridging large gaps in neural network representations with model stitching. In *Proceedings of UniReps: the Third Edition of the Workshop on Unifying Representations in Neural Models*, 2026.

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

Harshil Vejendla. Drift-adapter: A practical approach to near zero-downtime embedding model upgrades in vector databases. In *EMNLP*, pp. 15938–15949, 2025.

Krispin Wandel, Jingchuan Wang, and Hesheng Wang. ViT-Up: Faithful feature upsampling for vision transformers. *arXiv preprint arXiv:2606.14024*, 2026.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10506–10518, 2019.

Wen-Ting Wang, Wei-Ying Wu, Hao-Yun Huang, and Xuan-Chun Wang. Spatial adapter: Structured spatial decomposition and closed-form covariance for frozen predictors. *arXiv preprint arXiv:2605.11394*, 2026.

Di Wu and Xiaohui Zhu. Post-hoc sparse coding of latent communication between vision-language model agents. *arXiv preprint arXiv:2608.10198*, 2026.

Jinsung Yoon and Sercan O. Arik. Embedding-converter: A unified framework for cross-model embedding transformation. In *ACL*, pp. 25464–25482, 2025.

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

Binjie Zhang, Yixiao Ge, Yantao Shen, Shupeng Su, Fanzi Wu, Chun Yuan, Xuyuan Xu, Yexin Wang, and Ying Shan. Towards universal backward-compatible representation learning. In *IJCAI*, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

Qifeng Zhang, Ting Xiang, Zeyuan Bai, and Changjian Chen. Semantically compatible knowledge distillation for cross-domain object detection with vision foundation models. *arXiv preprint arXiv:2608.20916*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Dongfang Zhao. EGA: Adapting frozen encoders for vector search with bounded out-of-distribution degradation. *arXiv preprint arXiv:2605.05674*, 2026.

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive generation. In *NeurIPS*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026.

Yifei Zhou, Zilu Li, Abhinav Shrivastava, Hengshuang Zhao, Antonio Torralba, Taipeng Tian, and Ser-Nam Lim. BT^2: Backward-compatible training with basis transformation. In *ICCV*, 2023.

Xuanyu Zhu, Yan Bai, Yang Shi, Yihang Lou, Yuanxing Zhang, Jing Jin, and Yuan Zhou. Beyond the last layer: Multi-layer representation fusion for visual tokenization. *arXiv preprint arXiv:2605.10780*, 2026.

APPENDIX

Algorithm 1 CTFM: offline fitting and release-wide deployment.

Input: Frozen release models; declared responses and mode m (IQ-SAFE or QUALITY); stage-specific data partitions, update budgets, and checkpoint grids.
Output: A fixed map A_{release} and its base/anchor dependencies.

▷ *Offline targets and population transport*
Fix stage-specific fitting/selection sets and evaluation scope (App. B).
Freeze generator, encoder, and decoder weights. ▷ Gradients may pass through them.
for each field Z required by offline fitting or selection **do**
  Construct $x_\star(Z) = D_{\text{new}}(Z)$ and the declared response target $Y(Z)$.
  Cache targets as permitted by the host protocol. ▷ Targets share the migrated field.
end for
Fit and freeze standardization, g, and B on their prescribed calibration data.
Set $A_0(Z) = \mathbf{1}g(Q(Z)) + B(Z)$, with $Q(B(Z)) = 0$.
Set $(O, \Pi_m) = (Q, P_0)$ for IQ-SAFE, or $(0, I)$ for QUALITY.
▷ Ownership constrains only the correction around the calibrated base.
Initialize $h_\theta = 0$: rank-32 zero-output branch (RAEv2), paired rank-16 branches (PE).
for each prescribed compact-stage update, including the RAEv2 RGB warm-start **do**
  Form $A^m_\theta(Z) = A_0(Z) + \Pi_m h_\theta(Z)$ on the current minibatch.
  Evaluate the stage response: RGB then cycle (RAEv2), or decoder-final features (PE).
  Update only residual parameters with the host-specific loss (App. B).
  Record functional risk on the prescribed selection split at scheduled checkpoints.
end for
Retain the compact checkpoint under the host-specific fitting protocol.
Complete any prescribed selected-horizon refit; freeze the compact map.
Set A_{release} to the selected/refitted compact map.
▷ PE uses its selected horizon for refitting; see App. B.
if the scalable RAEv2 IQ-SAFE configuration is requested **then**
  Freeze $A_c \leftarrow A_{\text{release}}$; fit then freeze $V(Z) = ZW^\top + \mathbf{1}b^\top$.
  Compute calibration displacements $R_Z = P_0[V(Z) - A_c(Z)]$.
  for each prescribed rank/scale candidate (r, α) **do**
    Obtain U_r from leading eigenvectors of $\sum_Z R_Z^\top R_Z$.
    Set $C_r = W^\top U_r$ and construct A^0_{comp} using Eq. (8).
  end for
  Select (r, α) by the fixed labeled-development rule in App. A.5.
  Initialize $\theta_0 = (C_r, U_r)$; fix $s_k(Z) = e_k(Z; \theta_0)$ for each response k.
  ▷ Subsequent refinement and checkpoint selection use no class labels.
  for each prescribed factor-refinement update **do**
    Form $A_\theta(Z) = A_c(Z) + P_0\Delta_\theta(Z)$ with the shared U_θ factor.
    Update only (C_θ, U_θ) using Eq. (9).
    At scheduled checkpoints, record the fixed selection-set response risk.
  end for
  Select the minimum-risk refinement checkpoint $\theta^\star$.
  Freeze $A_{\text{release}}(Z) = A_c(Z) + P_0\Delta_{\theta^\star}(Z)$.
end if
Evaluate the frozen map on the reserved population, including ownership checks.
Export only map weights and required base/anchor state; target models remain offline.
▷ Deployment: one fixed field map; no per-input fitting or candidate search.
for each new field Z **do**
  Return $\widetilde{Z} = A_{\text{release}}(Z)$ to the installed consumers.
  Execute $D_{\text{old}}(\widetilde{Z})$ and, if installed, $\psi(Q(\widetilde{Z}))$.
end for

Release workflow. Alg. 1 gives the release workflow, including compact fitting and the optional scalable RAEv2 IQ-SAFE stage. Blue comments mark fitting, green ownership/deployment, purple selection, and gray fixed-model notes. The pseudocode separates labeled initialization selection from label-free response refinement; host-specific splits, checkpoint rules, and PE's selected-horizon refit remain specified in App. B. Only the returned field map and installed consumers execute at deployment.

Appendix roadmap. App. A gives formal details, tangent initialization, and response-guided refinement. App. B specifies both release graphs and optimization protocols. Comparator access and the VFM reproduction appear in App. C. Complete results follow in App. D, transfer in App. E, qualitative evidence in App. F, deployment accounting in App. G, and additional analyses in App. H.

Across tables, pale blue identifies CTFM, pale orange identifies an ownership-free alignment, and gray marks reference conditions. Effects are oriented so positive is favorable unless a caption states otherwise. Tables use point estimates in the main paper and retain uncertainty here.

A Proofs and formal details

The formal results distinguish coordinate invariance, constrained response loss, and functional coverage. Projection establishes exact preservation; a local smoothness bound characterizes the response cost of excluded coordinates; and functional coverage identifies downstream computations controlled by agreement with the fitted response. The compact and scalable implementations follow these results.

A.1 Geometry of quotient projection

Mean removal makes a correction invisible to mean-pooled consumers: Q averages token rows, and P_0 subtracts that average. By linearity, $Q(P_0H) = Q(H) - Q(\mathbf{1}Q(H)) = 0$. Applying Q to Eq. (3) with $\Pi_O = P_0$ gives $Q(A_\theta^{\text{IQ-SAFE}}(Z)) = g(Q(Z))$. Composition with any ψ proves invariance for consumers of the form $\psi \circ Q$. The implementation uses one compensated centering step around an otherwise float32 map. On the system evaluation sets, maximum pooled-coordinate error is below 10^{-6} for RAEv2 and 3.8×10^{-6} for PE.

For the norm identity, let $C = \mathbf{1}Q(R)$ and $F = P_0R$. Since $\mathbf{1}^\top F = 0$, $\langle C, F\rangle_F = 0$. Pythagoras gives $\|R\|_F^2 = \|C\|_F^2 + \|F\|_F^2$, and repeated rows yield $\|C\|_F^2 = n\|Q(R)\|_2^2$. Thus P_0R is the unique nearest correction in $\ker Q$. For a local update $H\delta$ and consumer Jacobian J,

$$\|J[\mathbf{1}Q(H\delta)]\|_2 \le \|J\|_{F\to 2}\sqrt{n}\|Q(H\delta)\|_2, \tag{10}$$

which bounds the first-order response removed by projection. Here J maps field perturbations to first-order consumer-response changes, and $\|J\|_{F\to 2}$ is the operator norm from field Frobenius norm to response ℓ_2 norm. This is the local response interpretation of the excluded energy in Eq. (6).

A.2 Local smooth contract-gap bound

Exact coordinate invariance does not imply an unchanged optimal response loss. The following bound separates the contribution of an excluded mean component from approximation error within the feasible correction space.

Proposition A.1 (smooth contract-gap bound). For a fixed input field, let $\mathcal{L}(R)$ be a differentiable response-surrogate risk as a function of the correction, and let R^* be an unrestricted minimizer with $\nabla\mathcal{L}(R^*) = 0$. Let $\mathcal{R}_{\text{IQ-SAFE}} \subseteq \ker Q$ be the safe actuator's reachable set and $\epsilon = \inf_{R\in\mathcal{R}_{\text{IQ-SAFE}}} \|R - P_0R^*\|_F$. If $\mathcal{L}$ is β-smooth along the relevant segments, then

$$0 \le \inf_{R\in\mathcal{R}_{\text{IQ-SAFE}}} \mathcal{L}(R) - \mathcal{L}(R^*) \le \frac{\beta}{2}\left(n\|Q(R^*)\|_2^2 + \epsilon^2\right). \tag{11}$$

The proposition requires smoothness only along the local segments connecting the unrestricted and ownership-feasible corrections. It separates the local surrogate bound into removed quotient energy and actuator reachability. The ownership interventions measure the quality–continuity trade-off directly, while the exact invariance identities hold independently of this local analysis.

Proof. For any $\widehat{R} \in \mathcal{R}_{\text{IQ-SAFE}}$, local smoothness and stationarity give

$$\mathcal{L}(\widehat{R}) \le \mathcal{L}(R^*) + \langle \nabla\mathcal{L}(R^*), \widehat{R} - R^*\rangle + \frac{\beta}{2}\|\widehat{R} - R^*\|_F^2 \tag{12}$$

$$= \mathcal{L}(R^*) + \frac{\beta}{2}\|\widehat{R} - R^*\|_F^2. \tag{13}$$

Both $\widehat{R}$ and P_0R^* are centered, so $\widehat{R}-P_0R^*$ is orthogonal to $P_0R^*-R^* = -\mathbf{1}Q(R^*)$. Consequently,

$$\|\widehat{R} - R^*\|_F^2 = \|\widehat{R} - P_0R^*\|_F^2 + n\|Q(R^*)\|_2^2. \tag{14}$$

Taking the infimum proves Eq. (11).

A.3 Functional coverage

Coordinate preservation is an algebraic identity, whereas response agreement controls consumers that factor through the declared responses with bounded sensitivity to perturbations. The following lemma states this coverage condition explicitly.

Lemma 1 (coverage of declared functionals). Let $F = (f_1, \dots, f_K)$ collect the functionals named by the release contract. For fixed weights $w_k > 0$, equip its response space with the norm $\|v\|_w = (\sum_{k=1}^K w_k\|v_k\|_2^2)^{1/2}$ and define

$$d_F(Z, Z') = \|F(Z) - F(Z')\|_w = \left(\sum_{k=1}^K w_k\|f_k(Z) - f_k(Z')\|_2^2\right)^{1/2}. \tag{15}$$

Here tensor-valued responses are vectorized. If $u = r \circ F$ and r is L-Lipschitz from this weighted response norm to the consumer's Euclidean norm, then

$$\|u(Z) - u(Z')\|_2 \le L d_F(Z, Z'). \tag{16}$$

This follows directly from Lipschitz continuity. The corresponding boundary for consumers outside the declared functional family is stated in App. H. For example, a released linear head $u(Z) = W \operatorname{vec} F(Z) + b$ has an explicit factorization through F in its computation graph. With an unweighted Euclidean response norm, $L = \|W\|_2$; fixed weighted norms can be absorbed into W. Its output error is therefore bounded by the response mismatch. For an argmax decision, a reference logit margin exceeding $2Ld_F(Z, Z')$ suffices to preserve the winning class. The released computation graph supplies the coverage condition for this head.

A.4 Host-specific compact actuators

RAEv2 uses a single rank-32 branch with a zero-initialized output matrix. PE uses two rank-16 branches sharing the token-mixing matrix; paired cancellation preserves the calibrated base at initialization while retaining nonzero parameter gradients: two nonzero branches have cancelling outputs. Let $\mathcal{N}(Z)$ denote channel-wise standardization by fixed host moments, $R = P_0\mathcal{N}(Z)$, and $q = Q(\mathcal{N}(Z))$. The PE compact actuator uses

$$h_j(R, q) = \mathrm{GELU}\big(T\big[(RD_j^\top) \odot \mathbf{1}\big(2\sigma(qG_j^\top)\big)\big]\big)\, U_j^\top, \qquad h_\theta(Z) = h_a(R, q) + h_b(R, q), \tag{17}$$

Here fields use the $n \times d$ row-token convention, $D_j, G_j \in \mathbb{R}^{r\times d}$, $U_j \in \mathbb{R}^{d\times r}$, and $T \in \mathbb{R}^{n\times n}$ with $r = 16$. The symbol $\odot$ denotes elementwise multiplication, σ the sigmoid, and GELU the activation. The gate $2\sigma(qG_j^\top)$ multiplies each bottleneck channel by a pooled-feature-dependent weight. Thus D_j, U_j form the channel bottleneck, T exchanges information across token positions, and $qG_j^\top$ conditions the residual on the pooled input. We initialize $D_b = D_a$, $G_b = G_a$, and $U_b = -U_a$. The output is therefore zero while both branches retain nonzero first-order gradients. For output gradient v, both output maps receive $v\phi^\top$; an optimizer step breaks their initial antisymmetry and gradients also propagate through D_j and G_j. RAEv2 uses the same branch expression with $r = 32$, unstandardized source coordinates, and $h_\theta = h_a$, initializing $U_a = 0$. Initially only its output matrix receives gradients; subsequent updates also train the input, gate, and token-mixing matrices. Both architectures contain 163,840 weights. The paired-initialization evidence in Tab. 31 concerns PE.

A.5 Ownership-feasible tangent initialization

The scalable construction augments the compact anchor with an additional correction to the centered affine displacement, preserving its protected pooled coordinate. Let A_c be the frozen IQ-SAFE anchor, let $V(Z) = ZW^\top + \mathbf{1}b^\top$ be a frozen affine aligner, and let $U_r \in \mathbb{R}^{d\times r}$ have orthonormal columns. Since P_0 removes constant token rows,

$$P_0V(Z) = (P_0Z)W^\top. \tag{18}$$

The rank-r projection of the centered affine displacement is

$$\alpha[P_0V(Z) - P_0A_c(Z)]U_rU_r^\top = \alpha\big[(P_0Z)(W^\top U_r) - (P_0A_c(Z))U_r\big]U_r^\top. \tag{19}$$

The operator $R \mapsto P_0RU_rU_r^\top$ is the orthogonal projector onto $\mathcal{T}_r = \{XU_r^\top : Q(X) = 0\}$. The factorized affine displacement stores U_r and $C_r = W^\top U_r$, proving the initialization identity in Eq. (8). A compensated centering step is applied before the correction is added to A_c.

To estimate the channel basis, we compute the centered calibration displacements $R_Z = P_0[V(Z) - A_c(Z)]$. The columns of U_r are the leading eigenvectors of the channel Gram matrix $\sum_Z R_Z^\top R_Z$, ordered by decreasing eigenvalue. They retain the channel directions with the largest displacement energy. This estimates a basis from existing affine outputs; it does not fit another input-dependent coefficient predictor.

For $d = 1024$ and rank 432, the two tangent matrices contain 884,736 scalars; together with the 163,840-parameter anchor, the map contains exactly 1,048,576 learned parameters. It is 1,024 parameters smaller than the full affine VFM map. The fixed grid comprises ranks $\{32, 64, 128, 192, 224, 256, 320, 384, 432\}$ and coefficients $\{0, 1/32, 1/16, 1/8, 1/4, 1/2, 3/4, 1\}$. The basis is estimated from 768 calibration states; 128 states from 32 disjoint development classes select the initialization. Admissibility requires finite outputs, pooled error at most 3×10^{-5}, positive lower bounds from class-level bootstrapping over the compact FFM-to-task control for both semantic endpoints, LPIPS and cycle, and PSNR non-inferiority with a -0.10 dB lower-bound margin. Among admissible candidates, the fixed lexicographic order favors minimum semantic gain, semantic-gain sum, LPIPS reduction, PSNR gain, lower rank, and lower coefficient. It selects rank 432 and $\alpha = 1$. This initialization selection uses labeled development semantics; subsequent response-refinement checkpoint selection uses the label-free objective in Eq. (9). Both matched refinement arms inherit this same selected initialization. App. H.9 reports the initialization rank profile.

In a development-only numerical check, the float32 factored implementation with compensated mean replacement agrees with its source formula to 4.77×10^{-7} maximum absolute error and 3.93×10^{-8} RMS, with pooled-coordinate error below 2.80×10^{-7}.

A.6 Response-guided refinement at fixed capacity

Response-guided refinement optimizes the two factors of the projected affine initialization while retaining the compact anchor. Its objective matches the archived decoder's RGB output and the archived encoder's features to their respective matched-native references. For RAEv2, define the two archived responses

$$F_{\text{old}}^{\text{rgb}}(X) = D_{\text{old}}(X), \qquad F_{\text{old}}^{\text{cyc}}(X) = E_{\text{old}}(D_{\text{old}}(X)), \tag{20}$$

with targets $Y^{\text{rgb}}(Z) = D_{\text{new}}(Z)$ and $Y^{\text{cyc}}(Z) = E_{\text{old}}(D_{\text{new}}(Z))$. The decoder uses the direct-RGB convention: its output is interpreted in image RGB units and clamped to $[0, 1]$, without a second image-space denormalization. For each response, the per-state squared error is divided by the error of the frozen compiled initialization on that state, with denominator floor 10^{-8}. The refinement minimizes Eq. (9) with AdamW, learning rate 10^{-5}, zero weight decay, batch size four, and 2,400 updates. Feature matching uses the same factorization, initialization, examples, order, optimizer, regularizer, checkpoint grid, and ownership projection. Only the fitted objective changes.

Dividing by the frozen initialization error compares relative progress across responses and inputs. For the selected two-response fit, initial image MSE denominators range from 0.003404 to 0.096230 and cycle denominators from 0.259729 to 1.504290. Thus no denominator reaches the 10^{-8} floor. At initialization, each active normalized response equals one up to numerical precision; subsequent minimax updates prioritize the response with the largest error relative to its own initialization value. Fixed per-state denominators retain each response's relative repair scale throughout fitting; the positive floor handles the zero-error case.

B Complete release and optimization specification

This section distinguishes reused upstream components from fitted migration maps. Tab. 5 summarizes the components, fixed transports, responses, and optimization rules for the two host systems.

Table 5: Release configurations and optimization rules. The functional and contract are fixed before the corresponding evaluation population is opened.

Host	State	Frozen transport	Declared functional
RAEv2	256×1024	ridge field transport followed by a quotient-conditioned base	compact: RGB then cycle; refinement: joint RGB and cycle
PE	256×1024	pooled-coordinate transport, centered-token transport, and incumbent residual	archived decoder-final 256×1152 hidden field

Host	Actuator	Optimization and selection
RAEv2	single rank-32 compact anchor; rank-432 tangent; 1,048,576 parameters after scaling	labeled development selects initialization; label-free normalized response risk selects factor refinement
PE	rank-16 paired actuator; 163,840 parameters	decoder-final response matching; minimum calibration functional risk

Migration assumes pretrained upstream models. The generated-field evaluations sample G_{new} directly; the real-image evaluations use $E_{\text{new}}(x)$. Both supply the same dense-field input type to the fitted map. The native decoder supplies offline references and is not substituted for the retained decoder in the migration deployment.

Released components and constructed pairings. The RAEv2 producer uses the released dinov3l-k7 ImageNet stage-2 checkpoint and matched decoder/statistics; the archived decoder/statistics are dinov2l-k1. The PE producer and matched decoder/statistics are the DiffusionBench spatialpe-vit-l release, paired with the released pe-l-k1 decoder and PE-Core-L encoder. These upstream generators, encoders, and decoders are not retrained in our migration experiments. We assemble the cross-release pairings, estimate the population transports, and fit the field maps. The supplement identifies the exact released assets. Matched-native targets are generated by executing the successor decoder on the same field; they require no manual cross-release image correspondence. The suffixes in the typewriter checkpoint names are retained upstream configuration identifiers; the old/new direction in the paper refers to the representation pair; the suffixes retain their upstream naming conventions.

Compact-stage objectives. RAEv2 first fits an RGB warm-start for 300 updates with AdamW, learning rate 10^{-3}, and batch size eight. Its loss is RGB MSE plus a 10^{-4} relative-displacement penalty from A_0. The ensuing cycle stage normalizes semantic-cycle MSE by the minibatch mean error of this frozen RGB warm-start:

$$\mathcal{L}^{\mathcal{B}}_{\text{compact}} = \frac{\bar{e}^{\mathcal{B}}_{\text{cyc}}(h_\theta)}{\max\{\bar{e}^{\mathcal{B}}_{\text{cyc}}(h_{\text{ref}}), 10^{-8}\}} + 10^{-4}\frac{\text{mse}_{\mathcal{B}}[A_\theta - A_0]}{\max\{\text{mse}_{\mathcal{B}}[A_0], 10^{-8}\}}. \tag{21}$$

Bars denote minibatch means; h_{ref} is the RGB warm-start. Cycle fitting uses AdamW (Loshchilov & Hutter, 2019), batch size four, learning rate 10^{-4}, and zero weight decay: an initial 300-update fit is followed by 3,600 continuation updates, restarting the optimizer after the first 1,200 continuation updates. The compact anchor is then frozen. During cycle fitting, RGB decoding supplies the differentiable encoder input without a separate RGB loss. Factor refinement jointly fits RGB and cycle (Eq. (9)). PE instead minimizes decoder-final feature MSE with a 10^{-4} relative-displacement penalty; its horizon selection and refit are specified below.

PE optimization. The PE calibration cache contains four states for each of 128 classes. A fixed 96/32 class split selects checkpoints every 50 updates, after which the selected horizon is refit on all 512 states. The 163,840-parameter CTFM arm uses AdamW, batch size two, learning rate 10^{-3}, zero weight decay, a 10^{-4} relative-displacement penalty, and at most 900 updates. Its target is the archived decoder-final hidden field and the earliest validation minimum breaks ties. FFM-to-TLT uses the same actuator and first fits that hidden field for 450 updates at 10^{-3}, then continues with gallery task loss at 10^{-4} through update 900; selection is among updates 450, 500, ..., 900 by validation task loss.

Tab. 6 distinguishes comparison roles, fitted stages, and evaluation populations. Generated samples with different seeds remain distinct even when their conditioning classes coincide.

Table 6: Experimental populations and fitted stages. All evaluation classes exclude the corresponding residual/refinement fitting classes. The 648-class source additionally excludes 32 compiler-selection and 128 earlier objective-selection classes (App. B).

Comparison	Compared maps	Residual/refinement fit	Evaluation population
RAEv2 objective	response / feature refinement	160 classes × 4 states each	808 classes × 4 new states
RAEv2 reduced data	40-class response / 160-class feature refinement	40×4 / 160×4 states	same 808 classes, separate new seeds
RAEv2 transfer	response / feature refinement	160 classes × 4 states each	648-class source and four ImageNet shifts
PE full data	compact CTFM / FFM-to-TLT	128 classes × 4 states each	800 classes × 4 states
PE reduced data	64-class CTFM / 128-class FFM-to-TLT	64×4 / 128×4 states	800 classes × 2 new states
RAEv2 illustration	compiled map / compact FFM-to-TLT	frozen maps	12 preselected states from the 648-class frontier

Reduced-data accounting boundary. The release transport g, B is fixed before every comparison. For reduced-data RAEv2, the compact anchor A_c, affine source V, tangent basis U_r, and compiled initialization A^0_{comp} are also shared and frozen; only response refinement uses 40 rather than 160 classes. For reduced-data PE, the transport and base are shared and frozen, while the trainable residual uses 64 rather than 128 classes. The reduced-data claims therefore concern the trainable residual/refinement stage. Data used to construct frozen release dependencies remain fixed across each comparison.

Class and sample independence. The RAEv2 base and compact stages use 192 ImageNet classes (Deng et al., 2009; Russakovsky et al., 2015), split into 160 fitting and 32 validation classes for factor refinement. A separate 32-class development set selects the compiler rank and coefficient; another 128 classes (512 states) supported earlier semantic-cycle objective selection. These three class sets are disjoint. The system population is their complement: $1000 - 192 - 32 - 128 = 648$ classes, with four states per class. Exclusions follow the recorded class-role assignments. The 808-class prospective populations exclude calibration classes and use fresh generation seeds after freezing the maps, but include both selection sets. Their 776-class complements exclude compiler-selection classes while retaining the 128 earlier objective-selection classes. Both full-data and quarter-data comparisons retain their original populations and separate seeds. The same scope applies to the regional and response-component subsets, optimizer-order repeats, fitting-breadth evaluations, and RAEv2 distribution summaries using these classes. The rank profile uses the 32-class compiler development set. PE's 800-class system and prospective populations exclude its 128 fitting classes and 72 additional development classes. Tab. 35 evaluates the compiler-selection exclusion, using class provenance alone, with no refitting or outcome-dependent filtering.

Cached-reference precision. RAEv2 caches native evaluation images as 8-bit RGB while retaining archived semantic encodings computed before image quantization. Pixel and perceptual scores use the cached image; cycle MSE uses the cached semantic target. PE caches native images and target fields separately in float16 and reloads them in float32. Re-encoding a cached native image need not reproduce its separately cached target exactly: cache rounding and numerical execution contribute to the reported native cycle floor (0.00321 for prospective RAEv2 and 0.0255 for the PE system population). All compared maps use the same references within each evaluation. Native self-comparison gives zero LPIPS and capped PSNR, while its cycle floor is retained rather than reset.

B.1 Frozen transport in each host system

The base A_0 is fitted before residual training. Its pooled output defines the protected coordinate, and its centered output supplies the initial spatial field. The host-specific transports implement this separation through mean replacement. Define compensated assembly as

$$\text{Comp}(X, q) = P_0 X + \mathbf{1}q, \tag{22}$$

implemented with compensated centering and a final mean replacement.

For RAEv2, the first stage is the fixed ridge transport

$$L(Z) = Z + \mu_r + (Z - \mu_x)W, \tag{23}$$

In this construction, L denotes the base transport and W its fitted channel matrix; both symbols are local to this definition. The vectors μ_x and μ_r are calibration means of the source fields and target-minus-source residuals. The matrix W minimizes paired field reconstruction under a calibration-selected ridge penalty. The base $A_0^{\mathrm{RAEv2}}(Z) = M_{\mathrm{QCM}}(Z, L(Z))$ uses a rank-32 quotient-conditioned morphism (QCM): it applies separate low-rank residuals to $Q(Z)$ and $P_0 Z$, conditions the centered residual on the quotient, re-centers it, and adds it to the transported field.

For PE, let $q = g_{\mathrm{PE}}(Q(Z))$ be the pooled-coordinate map. It standardizes the successor mean, rotates it into the shared calibration basis, applies the quotient corrector, restores its norm, rotates to the archived basis, and reverses archived standardization. With selected centered transport $T_{\mathrm{PE}}(Z) = (Z - \mu_x)R + \mu_y$ and incumbent residual I_{inc}, the PE base is

$$A_0^{\mathrm{PE}}(Z) = \mathrm{Comp}(I_{\mathrm{inc}}(Z, \mathrm{Comp}(T_{\mathrm{PE}}(Z), q)), q). \tag{24}$$

All moments, rotations, correctors, and incumbent weights are fixed before residual fitting and included in complete deployment accounting.

B.2 PE ARCHIVED FUNCTIONAL

The PE decoder exposes a hidden field before its final output layer. Matching that field provides a response target without requiring a task-label loss. To obtain the target in archived coordinates, we decode the successor field and re-encode its image with the archived encoder. Let H_{old} denote the archived decoder prefix through its final hidden token field. PE uses

$$F_{\mathrm{old}}^{\mathrm{PE}}(X) = H_{\mathrm{old}}(X), \qquad Y^{\mathrm{PE}}(Z) = H_{\mathrm{old}}(E_{\mathrm{old}}(D_{\mathrm{new}}(Z))). \tag{25}$$

This is final-feature matching under the ownership constraint. The FFM-to-TLT comparison evaluates the additional effect of the specified task-loss continuation; its schedule and checkpoint selector are stated above. Both migration maps deploy without the archived encoder. PE caches its old-encoded targets and fits through the frozen decoder prefix; RAEv2 evaluates its semantic-cycle response through the frozen archived encoder during offline refinement. Caching targets does not remove the differentiable consumer calls needed to evaluate migrated fields during fitting.

C COMPARISON SCOPE FOR PUBLISHED METHODS

Numerical comparison requires a common deployment interface. A direct field-map comparator must consume the same paired release states, leave the released generator, encoders, and archived decoder immutable, and return the same dense field. Tab. 7 categorizes recent pooled compatibility, frozen stitching, tokenizer redesign, generative adaptation, dense-feature transformation, task-specific interface recovery, and receiver-communication methods through the 30 August 2026 literature cutoff. Methods outside the declared release graph are compared by intervention rights and returned object; Tab. 8 then isolates direct field-map comparators from methods that alter update time, executable graph, or output. Direct numerical results compare the same output and evaluation target under the same deployment operations; the closest VFM objectives are reproduced at identical capacity and as a full affine frontier.

Post-release embedding conversion fits a map between released representations (Yoon & Arik, 2025; Vejendla, 2025; Hu et al., 2022b; Ricci et al., 2025). Training-time compatibility, orthogonal extensions, and backfilling operate at different stages of producer or archive maintenance (Shen et al., 2020; Zhou et al., 2023; Ricci et al., 2024; Jang & Lim, 2025; Seo et al., 2025; Bui et al., 2025). Backward-compatible image compression addresses evolving compression systems (Duan et al., 2024); CTFM instead fits a dense semantic field map after both releases are frozen. Frozen stitching connects network blocks; related single-consumer and receiver-communication methods return a stitched computation or a new message interface (Mai et al., 2026; Athanasiadis et al., 2026; Sanyal et al., 2026; Chaudhury, 2026; Li et al., 2026; Liu et al., 2026b; Wu & Zhu, 2026). Tokenizer, decoder, generative, and dense-task adapters instead modify a component or target a task-specific object (Chen et al., 2026; Liu et al., 2026a; Zhu et al., 2026; Huang et al., 2026; Zheng et al., 2025;

Ran et al., 2024; Mao et al., 2026; Jian et al., 2026; Go et al., 2026; Ouyang & Mao, 2025; Li et al., 2025; Wandel et al., 2026; Zhao, 2026; Wang et al., 2026; Zhang et al., 2026). Feasible-by-design projection layers also enforce hard constraints on neural outputs (Grontas et al., 2026); CTFM instead uses an analytic projector on a dense-field correction whose protected coordinates are declared by a consumer contract.

Table 7: Compatibility capabilities of published method families through 30 August 2026. Capacity matching isolates objective effects; the separately reported affine VFM frontier provides a strong system-level comparator.

Approach family	When	What changes	Returned interface
Embedding conversion	post-release	vector map or stored archive	pooled embedding or updated index
Backward-compatible training	successor training	representation producer	compatible embedding
Frozen model stitching	post-hoc	connection between network blocks	stitched computation
Tokenizer or decoder adaptation	training or post-training	encoder, decoder, or executable graph	revised latent or image
Dense task adaptation	training or post-hoc	task-specific feature adapter	task-oriented dense feature
CTFM	post-release	one field map	**archived-consumer-executable dense field**

Table 8: Comparator access and returned object. A check denotes an unchanged released component, *changed* denotes an intervention on that component, and a dash denotes a component outside the method's graph. "Addl. archived call by map" means an extra archived-model evaluation inside the migration map, before installed consumers execute its output.

	Fit access		Released components fixed			Deployment	
Intervention	**State pairs**	**Labels**	**Gen.**	**Enc.**	**Dec.**	**Addl. archived call by map**	**Return**
Frozen calibrated transport	yes	no	✓	✓	✓	no	field
Feature matching	yes	no	✓	✓	✓	no	field
FFM-to-TLT	yes	yes	✓	✓	✓	no	field
CTFM refinement	yes	no	✓	✓	✓	no	field
Decoder adaptation	yes	optional	✓	✓	changed	no	decoder/image
Training-time BCT	train set	yes	–	changed	–	no	embedding
Archive backfill	archive	optional	–	✓	–	no	updated archive
Exact/native references	native	no	✓	✓	native	offline	field/image

C.1 VFM OBJECTIVE REPRODUCTION

The label-access entries in Tab. 8 for feature matching and CTFM describe their fitting objectives. For scalable RAEv2, labeled development data select the *shared initialization*; both refinement arms inherit that initialization and subsequently fit and select checkpoints without class labels. FFM-to-TLT additionally uses labels in its task-loss stage.

The closest published objectives admit direct comparison within the required field interface when their fitting procedures are specified. The public VFM repository available at the 30 August 2026 comparison cutoff did not include an executable training implementation. We therefore reproduce its published matching objectives in the present release graph and describe the comparison as an objective-level reproduction. Our 144 cross-release fits comprise two hosts, three map families, four objectives, three learning rates, and two weight decays. Every candidate uses the same class-disjoint validation split, and the selectors are fixed before held-out evaluation. Full affine VFM and the identical-parameter rank-432 feature-matching actuator are both retained, so the comparison covers the strongest system frontier and the controlled objective contrast. The selected affine values are reported in the complete RAEv2 and PE system frontiers (Tabs. 2 and 33), where their ownership

status and the reference conditions remain visible in context. Tab. 9 enumerates the complete published-objective reproduction grid underlying that comparison.

Table 9: Exact objective-reproduction grid. The Cartesian product contains $2 \times 3 \times 4 \times 3 \times 2 = 144$ cross-release fits. Self-stitches are trained separately for every promoted configuration.

Axis	Fixed values
Host	RAEv2; PE
Map family	affine; parameter-matched two-layer ReLU MLP; width-$4d$ MLP
Objective	LFM; FFM; TLT; FFM-to-TLT
Learning rate	10^{-3}; 5×10^{-3}; 10^{-2}
AdamW weight decay	0; 0.01
Schedule	at most 100 epochs; patience 5; earliest validation minimum
Effective batch	256 for LFM; 128 otherwise
Selector	lexicographic: max–min semantics; semantic sum; LPIPS; PSNR; parameters; latency

D Complete quantitative results

The main paper reports compact point estimates. This section gives the full prospective comparison, its class-cluster uncertainty, residual-stage data efficiency, and distributional metrics. All RAEv2 values use the verified direct-RGB decoder contract.

D.1 Prospective equal-capacity confirmation

Prospective evaluation applies frozen refinement checkpoints to newly generated samples, testing the objective comparison beyond the earlier system population. The table distinguishes the identical-capacity control, full affine comparator, and reference conditions.

Table 10: **Absolute prospective RAEv2 results** on 3,232 states from 808 refinement-disjoint classes (App. B). EfficientNetV2-S (Eff.) and ViT-B/16 were absent from fitting and selection. Rows are grouped by comparison role; blue marks CTFM and gray marks references. Map weights include the retained anchor and exclude the shared frozen base; identical-image PSNR is capped at 120 dB.

Method	Map weights	Gallery ↑	ResNet-50 ↑	Eff. ↑	ViT ↑	PSNR ↑	LPIPS ↓	Cycle ↓
Feature matching	1,048.6K	52.970	72.246	75.124	73.855	15.885	0.29071	0.65059
CTFM	1,048.6K	**54.548**	**75.650**	**77.908**	**77.073**	**17.744**	**0.23830**	**0.52842**
Projected affine VFM	1,049.6K	53.806	**75.897**	77.537	76.423	17.196	0.24115	0.56272
Archived field	–	56.312	82.085	84.220	82.085	19.407	0.14448	0.29655
Matched-native image	–	57.364	81.528	83.540	81.993	120.000	0.00000	0.00321

Tab. 10 evaluates frozen maps on 3,232 new states from 808 classes excluding the maps' 160 refinement-fit and 32 refinement-validation classes. The upstream selection history and complementary 776-class analysis are specified in App. B and Tab. 35. The generator seeds are new and were absent from checkpoint selection and the earlier system frontier. CTFM and feature matching share the ownership-feasible rank-432 actuator, 1,048,576 parameters, initialization, examples, optimizer, updates, and checkpoint grid; only the objective differs. CTFM improves all seven endpoints against the identical-capacity control and six of seven against the 1,049,600-parameter affine VFM map. Tab. 11 reports the corresponding class-bootstrap effect estimates. Each bootstrap draw treats classes as the resampling units and retains all within-class samples together.

Table 11: **Class-bootstrap uncertainty for the prospective objective control.** Effects are CTFM minus identical-capacity feature matching; LPIPS and cycle entries are reductions, so every positive value favors CTFM. Intervals use 20,000 class-cluster bootstrap draws.

Endpoint	Effect	95% interval	Endpoint	Effect	95% interval
Gallery Top-1	+1.578 pp	[+0.619,+2.506]	PSNR	+1.860 dB	[+1.798,+1.922]
ResNet-50 Top-1	+3.403 pp	[+2.351,+4.455]	LPIPS reduction	0.05242	[0.05070,0.05421]
Independent Top-1	+3.001 pp	[+2.135,+3.868]	Cycle reduction	0.12217	[0.11876,0.12560]

D.2 PE OWNERSHIP-PRESERVING FRONTIER

Tab. 12 tests the same ownership requirement on the independently constructed PE host. The calibrated base is applied to the successor field before residual repair. CTFM is best on every endpoint among immutable field maps and also exceeds the graph-changing controls. The complementary ownership-free intervention is analyzed in App. H.10.

Table 12: **PE ownership-preserving system frontier** on 3,200 states from 800 held-out classes. Rows are grouped by deployment contract. Blue marks CTFM, and gray marks references. Maximum drift is measured from the protected pooled coordinate.

Method	Gallery ↑	ResNet-50 ↑	PSNR ↑	LPIPS ↓	Cycle ↓	Max drift ↓
Ownership-preserving field maps						
Calibrated base	18.38	43.34	14.510	0.4535	19.762	0.000
Complete-cycle map	20.88	50.44	14.604	0.4382	19.127	3.70×10^{-6}
Global task-loss map	21.47	49.09	14.624	0.4409	19.251	3.62×10^{-6}
FFM-to-TLT	24.81	45.38	17.027	0.4943	18.724	3.58×10^{-6}
IQ-SAFE CTFM	**26.56**	**54.44**	**17.200**	**0.4226**	**18.447**	3.70×10^{-6}
Graph-changing controls						
Pixel decoder LoRA	15.91	17.00	15.913	0.5519	20.034	–
Semantic decoder LoRA	18.16	39.69	14.450	0.5638	19.607	–
Relaxed decoder head	17.66	11.66	12.285	1.1286	19.498	–
Reference conditions						
Archived field	51.72	84.13	19.025	0.1767	10.797	–
Matched-native image	57.09	85.78	120.000	0.0000	0.0255	–

D.3 REGIONAL OWNERSHIP WITHOUT REFITTING

An installed consumer may store spatial descriptors instead of a single token mean. Let $\mathcal{S}$ partition the 16×16 token grid into four or sixteen equal rectangular regions, and let $S(i)$ be the region containing token i. The ownership operator collects their channelwise means, and its kernel projector subtracts the corresponding regional mean:

$$O_{\mathcal{S}}(Z) = \left(|S|^{-1} \sum_{i \in S} Z_i \right)_{S \in \mathcal{S}}, \qquad [\Pi_{\mathcal{S}} R]_i = R_i - |S(i)|^{-1} \sum_{j \in S(i)} R_j. \tag{26}$$

For a frozen candidate map A and the common retained anchor A_c, deploy $\widehat{A}_{\mathcal{S}}(Z) = A_c(Z) + \Pi_{\mathcal{S}}[A(Z) - A_c(Z)]$. Then $O_{\mathcal{S}}(\widehat{A}_{\mathcal{S}}(Z)) = O_{\mathcal{S}}(A_c(Z))$; because the regions partition all tokens, global mean ownership also follows. This is composition of a fitted map with a declared projector. All candidate maps and the anchor remain frozen, and the projector adds zero learned weights.

Tab. 13 reuses the first manifest state from each of the 808 refinement-disjoint classes in Tab. 10, chosen by fixed order before this comparison. Both rank-432 refinements use their original 2,400-update checkpoints. The affine comparator applies the same ownership operator relative to the same anchor. Every output is decoded and evaluated under the established direct-RGB protocol. Thus the new comparison changes the owned coordinates while holding the evaluated states and comparator access fixed.

Table 13: **A frozen map supports regional ownership.** Recognition is in percent; PSNR is in dB. Eff./ViT are the independent EfficientNetV2-S and ViT-B/16 models. Drift is the maximum absolute error in the declared means relative to the common anchor. Map weights include the anchor and exclude shared population transport. Bold marks the best task score within each three-map regional block.

Method	Weights	Gallery↑	ResNet↑	Eff.↑	ViT↑	PSNR↑	LPIPS↓	Cycle↓	Drift↓
Reference: global mean ownership									
Frozen anchor	163.8K	37.13	35.64	40.10	39.60	16.468	0.5260	1.1045	0.00
Feature matching	1,048.6K	52.10	72.15	74.38	74.13	15.808	0.2915	0.6465	3.35×10^{-7}
CTFM	1,048.6K	54.08	75.99	78.59	76.24	17.725	0.2373	0.5232	3.52×10^{-7}
Four regional means: 2×2 spatial partition									
Feature matching	1,048.6K	52.10	70.54	73.76	73.14	15.850	0.3027	0.6736	5.51×10^{-7}
Affine VFM	1,213.4K	51.49	75.74	**77.23**	75.74	17.175	0.2485	0.5796	6.26×10^{-7}
CTFM	1,048.6K	**53.22**	**75.87**	**77.23**	**76.49**	**17.661**	**0.2466**	**0.5457**	5.48×10^{-7}
Sixteen regional means: 4×4 spatial partition									
Feature matching	1,048.6K	50.50	67.33	71.91	68.81	16.079	0.3308	0.7370	5.51×10^{-7}
Affine VFM	1,213.4K	51.98	**73.64**	**76.49**	71.53	17.135	0.2722	0.6327	6.71×10^{-7}
CTFM	1,048.6K	**52.85**	73.27	76.11	**74.75**	**17.493**	**0.2717**	**0.6008**	1.03×10^{-6}

CTFM improves all seven reported endpoints over the identically parameterized feature-matching map under both regional contracts. Four-region protection retains 99.69% of the global map's ResNet gain above the anchor; sixteen-region protection retains 93.25%. The corresponding mean removed correction energies are 7.75% and 23.85%, with mean regional RMS errors of 2.40×10^{-8} and 3.11×10^{-8}. The unprojected global-mean CTFM changes these regional descriptors by as much as 1.91 and 2.96, respectively, so the experiment tests additional ownership requirements. Correction energy and recognition gain are not proportional: regional projection removes directions according to coordinate ownership, whereas the decoder's sensitivity determines their downstream effect. The energy measures removed field variation. Parameter storage is accounted for separately: the complete regional affine map stores 1,213,440 weights: 1,049,600 affine weights and bias plus the 163,840-weight anchor needed to supply the protected regional means. Both learned refinements store 1,048,576 weights including that anchor. App. H.4 supplies paired uncertainty and the individual response-component endpoints.

D.4 Residual-stage data efficiency

Residual-stage data efficiency is evaluated with previously fitted dependencies fixed; only correction-fitting data are reduced. Tab. 14 compares each limited-data CTFM residual stage with the full-data, identical-capacity control. RAEv2 uses four states per refinement class; PE uses its fixed host-specific protocol. Consumer-functional supervision improves every reported endpoint with one quarter of the RAEv2 response-refinement states or one half of the PE residual-fitting classes. The shared release dependencies are listed in Tab. 6. Complete breadth trajectories are retained as a sensitivity analysis in App. H.5.

Table 14: **Cross-host residual-stage data efficiency.** Entries are directed effects of limited-data CTFM over the full-data, identical-capacity control. LPIPS and cycle MSE are reductions, so every positive entry favors CTFM.

Host	CTFM data	Gallery	ResNet-50	Independent	PSNR	LPIPS red.	Cycle red.
RAEv2	25%	**+1.640**	**+3.868**	**+2.738**	**+1.823**	**0.04588**	**0.10420**
PE	50%	**+2.062**	**+10.875**	**+9.844**	**+0.125**	**0.09058**	**0.31390**

D.5 Cross-host distributional quality

Paired image errors assess correspondence to individual reference images; distribution metrics characterize the output collection. We use a fixed Inception feature extractor, writing $f(x) \in \mathbb{R}^p$ for an image's feature vector, with $p = 2048$. Native FID compares the feature means and covariances of migrated images and their matched-native references; real FID uses class-matched ImageNetV2

MatchedFrequency images as the reference. KID measures the corresponding feature-distribution discrepancy with an unbiased degree-three polynomial-kernel estimator.

The additional distances have a classwise interpretation. Paired feature MSE averages $\|f(x) - f(x_{\text{native}})\|_2^2/p$ for corresponding images. Class-centroid MSE compares the average feature vectors within each class. For generated and real feature sets $\mathcal{G}_c$ and $\mathcal{R}_c$ of class c, symmetric real adequacy is

$$d_{\text{adequacy}}(c) = \frac{1}{2p}\left[\frac{1}{|\mathcal{G}_c|}\sum_{u\in\mathcal{G}_c}\min_{v\in\mathcal{R}_c}\|u-v\|_2^2 + \frac{1}{|\mathcal{R}_c|}\sum_{v\in\mathcal{R}_c}\min_{u\in\mathcal{G}_c}\|v-u\|_2^2\right]. \tag{27}$$

The two directions measure proximity of generated features to real examples and coverage of real features by generated examples. We average these classwise distances equally over classes. All are measured in feature space, with lower values indicating better distributional agreement.

The distribution study applies the limited-residual-data CTFM maps from the strict cross-breadth comparisons to the complete prospective populations. Tab. 15 reports global FID and representative classwise feature distances. CTFM improves both FID references on both host systems and every primary classwise distance. These gains complement the paired PSNR, LPIPS, and semantic-cycle improvements; the complete KID sensitivity is retained in App. H.6.

Table 15: **Distributional quality after limited residual-stage fitting.** The baseline uses full refinement/residual-fit data; CTFM uses one quarter of the RAEv2 refinement states and half of the PE residual-fit classes. All metrics are risks, so lower is better. Native metrics use matched realizations; real metrics use class-matched ImageNetV2 MatchedFrequency images.

Host	Method	Native FID	Paired MSE	Real FID	Adequacy
RAEv2	Feature matching / full	20.2395	0.050885	20.0387	0.092381
	CTFM / reduced	**15.4744**	**0.040945**	**19.3026**	**0.088616**
PE	FFM-to-TLT / full	89.2179	0.123169	65.2921	0.136443
	CTFM / reduced	**74.9794**	**0.105432**	**55.1153**	**0.126148**

D.6 Joint response supervision and normalization

Per-state normalization is applied separately to each generated sample. Eq. (9) normalizes each response by its initialization error and minimizes the larger relative error on each input. We test its components at exactly 800 updates using the original seed, 640 fitting states, rank-432 initialization, minibatch order, AdamW settings, and displacement penalty. The three additional fits use RGB only, cycle only, or both responses with fixed population-average initial-error denominators. The existing joint per-state and feature-matching trajectories supply their 800-update checkpoints. All conditions use one seed and are evaluated on the same 808 samples as Tab. 13, one per class in fixed dataset order, without selecting a checkpoint on evaluation outcomes.

For interpretation, we analyze the data term of the declared objective on the fixed outputs. Let $\rho_k(Z) = e_k(Z;\theta)/\max\{s_k(Z), 10^{-8}\}$. Tab. 16 reports the mean and upper tail of $\max_k \rho_k$, together with the number of inputs for which both $\rho_{\text{rgb}} < 1$ and $\rho_{\text{cyc}} < 1$. These measurements describe which responses improve together on the fixed evaluation samples. RGB ratios are reconstructed from paired PSNR, cycle ratios from recorded MSE; the displacement regularizer is excluded.

Table 16: **Per-state supervision makes simultaneous repair widespread.** All fitted rows use 800 updates and 1,048,576 stored weights. Each response ratio is relative to the same initialization, whose worst-response ratio is one. Simultaneous repair requires strict improvement in both RGB and cycle MSE for the same input. Lower worst-response ratios and higher repair counts are better. Bold identifies the best fitted row in each column.

Objective	Mean worst ratio↓	95th percentile↓	Both repaired↑	Fraction↑
Feature matching	1.24549	1.78836	79/808	9.78%
RGB only	1.03692	1.17559	248/808	30.69%
Cycle only	0.99860	1.09850	416/808	51.49%
Joint, population normalization	0.96892	1.08375	640/808	79.21%
Joint, per-state normalization	**0.94963**	**1.02340**	**732/808**	**90.59%**

Per-state normalization gives the lowest mean worst-response ratio and the highest simultaneous-repair rate. Its gain over population normalization is 11.39 percentage points, with a paired 95% interval of [8.91,13.86]. This connects the loss design to its intended behavior: repair both declared responses across inputs. Tab. 26 gives every individual endpoint at the same horizon.

E TRANSFER AND EXTERNAL POPULATIONS

E.1 IMAGENET SHIFTS WITHOUT REFITTING

Frozen RAEv2 maps transfer to real-image fields from ImageNet-R (Hendrycks et al., 2021a), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and gallery-disjoint ImageNetV2-D (Recht et al., 2019). V2-D comprises the deduplicated union of Threshold0.7 (a-44) and TopImages (c-12), excluding every MatchedFrequency (b-33) gallery image by source identifier and file/decoded-RGB hashes. All 10,683 eligible images across 1,000 classes are retained, independently of model outputs. Dataset scores average class means equally; the macro averages the five populations equally. Tab. 18 reports effects against identical-capacity feature matching: all 25 standard and ten independent-observer cells favor CTFM. The macro also favors CTFM in all six endpoint families against projected affine VFM, whose map has 1,024 additional parameters.

Real-image field and metric references. Each manifest image is converted to RGB and bicubically resized to 256×256. RAEv2 comparisons use the same $[0, 1]$-clamped resized query for every map, both before encoding and as the reconstruction reference. The frozen successor encoder produces $Z = E_{\text{new}}(x)$ and the archived encoder produces the paired field $E_{\text{old}}(x)$. PSNR and LPIPS compare $D_{\text{old}}(A(Z))$ with the same real image x; cycle MSE compares $E_{\text{old}}(D_{\text{old}}(A(Z)))$ with $E_{\text{old}}(x)$. The shift evaluations therefore use the real input as their reconstruction reference.

Tab. 17 evaluates two computations on these same queries. The retrieval system reads the translated field's mean directly; the decoder uses the entire field to reconstruct the real image. Retrieval uses one reference centroid per class, so a correct result identifies the image's class through its gallery centroid.

Table 17: **Frozen pooled-consumer execution on disjoint real-image queries.** The archived gallery contains one DINOv2 centroid for each of 1,000 ImageNet classes, estimated from ten ImageNetV2 MatchedFrequency images per class before the evaluated maps. Query images share no gallery image. R@1, R@5, and mAP use class relevance; with one relevant centroid, per-query AP is reciprocal rank. Retrieval columns are percentages. PSNR and LPIPS evaluate the archived decoder on the same queries.

Queries	Field presented to consumer	R@1	R@5	mAP	PSNR	LPIPS
ImageNet-R	Direct reuse	0.06	0.37	0.68	8.542	0.8115
	Calibrated base	4.53	13.28	9.81	9.873	0.6752
	Projected affine VFM	4.53	13.28	9.81	14.828	0.2923
	IQ-SAFE CTFM	4.53	13.28	9.81	15.542	0.2833
	Unprojected affine VFM	0.25	0.59	0.99	16.871	0.2526
	Archived field	25.84	49.00	36.86	17.702	0.1692
ImageNet-Sketch	Direct reuse	0.15	0.55	0.76	7.932	0.8053
	Calibrated base	5.90	17.13	12.29	7.274	0.7135
	Projected affine VFM	5.90	17.13	12.29	11.134	0.3447
	IQ-SAFE CTFM	5.90	17.13	12.29	13.503	0.2913
	Unprojected affine VFM	0.15	0.68	0.96	17.608	0.2316
	Archived field	34.45	60.87	46.70	17.887	0.1436

The calibrated base, IQ-SAFE CTFM, and projected affine VFM yield identical reported retrieval metrics; scores agree to numerical precision. Across both domains, CTFM changes zero top-ranked decisions relative to the base and its maximum pooled-coordinate deviation is 4.77×10^{-7}. The table therefore separates two effects: base calibration recovers part of archived retrieval, and CTFM improves decoding without changing the calibrated query.

Tab. 18 then reports the complete zero-refit effect vector for the source and four shift populations, including V2-D.

Table 18: **RAEv2 transfer without refitting.** Effects compare response and feature refinements (Tab. 6). Recognition gains are in percentage points and PSNR gains in dB; LPIPS/cycle entries are reductions. Gallery is decoded-gallery Top-1; Eff./ViT are independent EfficientNetV2-S/ViT-B/16 Top-1. Positive entries favor CTFM.

Population	Gallery	ResNet-50	PSNR	LPIPS red.	Cycle red.	Eff.	ViT
Source generated (648 cls.)	+1.235	+4.128	+1.877	0.05513	0.12537	+3.665	+3.395
ImageNet-R	+1.250	+4.656	+1.835	0.05810	0.10077	+3.500	+5.187
ImageNet-Sketch	+1.350	+5.400	+3.700	0.14755	0.14304	+5.050	+5.575
ImageNet-A	+1.451	+1.563	+1.408	0.03948	0.13104	+1.136	+2.560
ImageNetV2-D	+1.682	+4.444	+1.638	0.04499	0.12357	+3.678	+4.878
Five-population macro	**+1.394**	**+4.038**	**+2.092**	**0.06905**	**0.12476**	**+3.406**	**+4.319**

Tab. 19 aggregates these five populations against the full projected affine VFM frontier.

Table 19: **Five-population macro effects against projected affine VFM.** CTFM has 1,048,576 parameters and VFM has 1,049,600. All directed class-stratified 95% intervals are favorable. Both maps use identical query preprocessing on each shift population.

Endpoint family	Effect	95% interval
Gallery Top-1	+0.395 pp	[+0.085,+0.710]
ResNet-50 Top-1	+0.670 pp	[+0.284,+1.055]
Independent Top-1	+0.710 pp	[+0.420,+0.995]
PSNR	+0.922 dB	[+0.901,+0.943]
LPIPS reduction	0.01541	[0.01480,0.01602]
Cycle reduction	0.03028	[0.02934,0.03122]

The independently constructed PE map shows the same zero-refit behavior. Relative to its matched FFM-to-TLT control, IQ-SAFE CTFM improves Gallery/ResNet-50 by $+0.44/+3.38$ points on ImageNet-R and $+0.98/+3.93$ points on ImageNet-Sketch. PSNR improves by 0.90 and 1.86 dB, LPIPS falls by 0.08361 and 0.09080, and cycle MSE falls by 0.33406 and 0.39794, respectively.

Class-outcome distribution. Fig. 3 complements dataset means by showing the distribution of class-level changes. A recognition tie denotes equal class-level accuracy, which can coexist with different individual predictions. The source evaluation has only four states per class. The continuous metrics are class means before distribution summaries are computed. For population $\mathcal{C}$ and class effects δ_c, the empirical CDF is $\widehat{F}(t) = |\mathcal{C}|^{-1} \sum_{c \in \mathcal{C}} \mathbf{1}\{\delta_c \le t\}$. The figure linearly connects its values at ordered class effects. Positive effects mean CTFM minus feature matching for PSNR and feature matching minus CTFM for LPIPS. No classes or tails are dropped.

E.2 TRANSFER TO INDUSTRIAL IMAGERY

Industrial transfer applies the frozen PE map to all 1,568 test images from the five MVTec LOCO categories (Bergmann et al., 2022), without domain-specific fitting or selection. Tab. 20 shows a consistent decoded-image advantage over the calibrated base and the matched FFM-to-TLT control.

Table 20: **Industrial transfer without refitting** on MVTec LOCO. Bold marks the best immutable learned field; gray rows are references.

Method	PSNR ↑	LPIPS ↓	Cycle ↓
Calibrated base	10.295	0.5098	21.158
FFM-to-TLT	14.584	0.3872	19.313
IQ-SAFE CTFM	**14.715**	**0.3347**	**19.162**
Archived field	16.722	0.2080	13.959
Matched-native image	20.225	0.1238	12.534

F QUALITATIVE EVIDENCE

Qualitative evaluation complements full-population metrics with localized comparisons of decoded outputs. The main illustration isolates identical-capacity refinement; the appendix includes paired error maps, cross-capacity initialization, PE, and real-image transfer. Each subsection specifies its models, references, and example-selection rule.

Image rendering. All examples use the direct-RGB decoder convention. We undo field normalization before decoding, clamp the decoder's RGB output to $[0, 1]$, and round the common 8-bit display conversion. No second image-space denormalization, sharpening, color correction, or per-method enhancement is applied. All visible color, texture, and sharpness differences are therefore properties of the decoded outputs.

Core-figure measurements. Fig. 1 uses held-out hair-slide state 1538: panel (a) supplies the matched-native reference, and panel (b) compares direct reuse, projected affine VFM, and the response-refined CTFM map. Its CTFM schematic summarizes compact repair; Fig. 2 expands the optional scalable construction used for the displayed output. The lower-left supervision paths in Fig. 2 use Pekinese state 392 for the native/CTFM images. Field tiles, point clouds, correction arrows, and response feature strips are schematic and encode no measured dimensions or activations. Photographs derive from the audited direct-RGB outputs. Selection is illustrative; quantitative results retain the full population.

Matched-capacity main illustration. Fig. 4 uses the response and feature refinements at the fixed 2,400-update checkpoint. Both maps contain 1,048,576 weights, including the retained anchor, and each refinement optimizes 884,736 factor weights. We applied a deterministic display rule to the complete fixed 2,592-state population: CTFM must be Gallery- and ResNet-50-correct, feature matching must miss at least one decision, PSNR gain must be at least 4 dB, LPIPS reduction at least 0.10, and cycle-MSE reduction greater than 0.05. Nine states qualify. They are ordered by PSNR gain plus twenty times LPIPS reduction plus ten times cycle-MSE reduction, with a stable population-index tie break; the main panel shows the first six from distinct classes. This outcome-aware rule is used only for illustration; Fig. 3 and all quantitative tables use their complete stated evaluation populations.

All displayed outputs use the same frozen states, checkpoints, and direct-RGB convention. The supplementary reproduction record documents rendering checks. Fig. 5 localizes the corresponding RGB reductions for the six main-text comparisons.

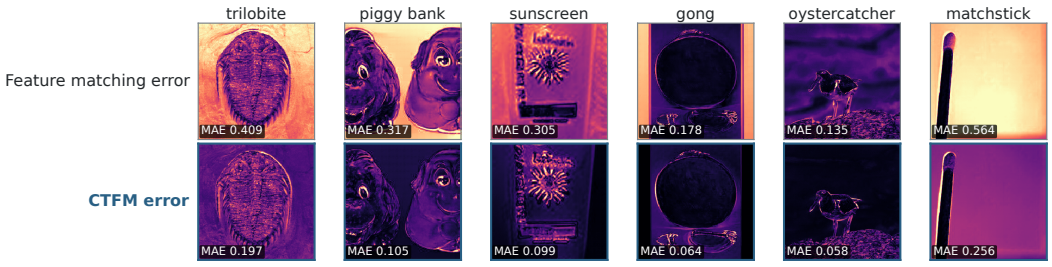


Figure 5: **Matched-capacity RGB error maps corresponding to Fig. 4.** Every panel uses the same display scale. The insets report mean absolute RGB error to the matched-native realization; all six CTFM maps substantially reduce MAE relative to feature matching.

Cross-capacity compiler illustration. For each sample in the 2,592-sample RAEv2 system evaluation, an incompatibility score quantifies initial failure under direct reuse as the mean of three empirical ranks: low PSNR, high LPIPS, and high semantic-cycle MSE relative to the matched-native image. The renderer first selects the 12 highest-scoring states from distinct classes without reading any learned-map outcome. From these, Fig. 6 retains the five cases where matched-native and CTFM are correct under both Gallery and ResNet-50, the compact FFM-to-TLT control misses at least one decision, and CTFM improves both PSNR and LPIPS. Eligible cases are ranked by the sum of four stable ranks: high absolute CTFM PSNR, low absolute CTFM LPIPS, high PSNR gain, and high

LPIPS reduction. This outcome-aware second stage is used only for illustration; all quantitative claims use the complete held-out populations.

The cross-capacity illustration compares the closed-form CTFM compiler (1,048,576 weights) with compact FFM-to-TLT (163,840), distinct from the matched refinement in Fig. 4. Both share the calibrated base and archived decoder.

Fig. 6 presents these five cross-capacity examples with the corresponding RGB error maps.

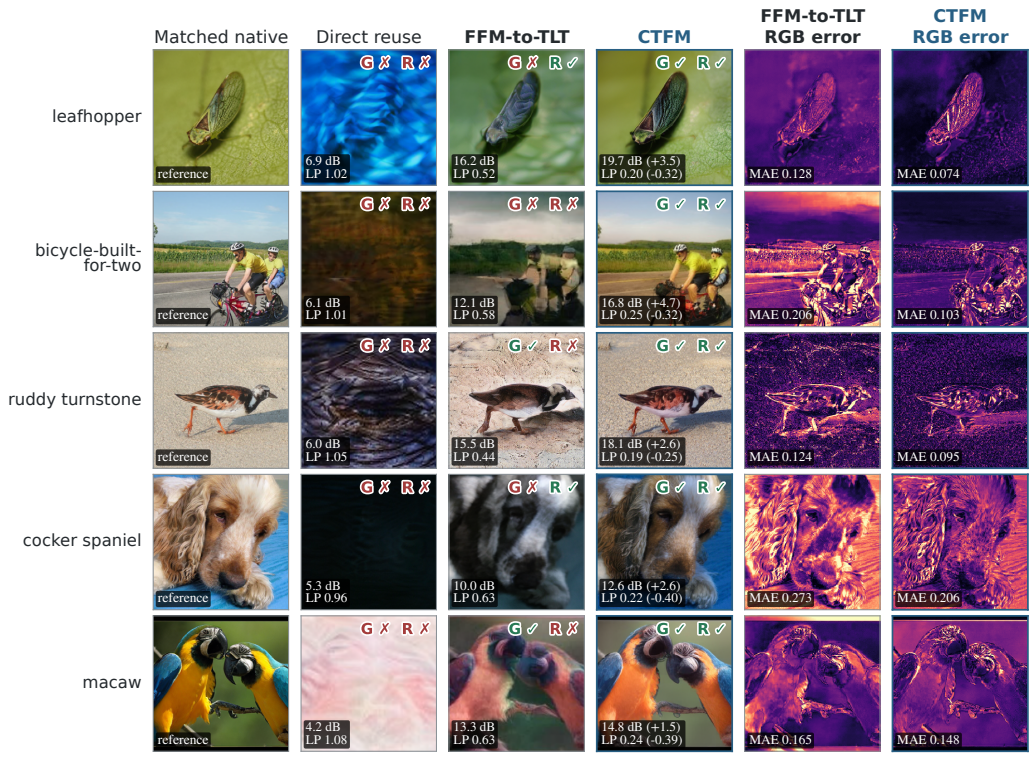


Figure 6: **Cross-capacity system illustration.** The closed-form CTFM compiler contains 1,048,576 map weights; the compact FFM-to-TLT control contains 163,840. G/R denotes Gallery/ResNet-50; ✓/× indicates a correct/incorrect target-class prediction. In-panel labels give PSNR/LPIPS and CTFM changes; both RGB-error columns share one scale and report MAE. These examples show the compiled initialization at larger capacity; the main text separately isolates the refinement objective at identical capacity.

Fig. 7 shows the complete set of PE generated states that meet the joint semantic-and-perceptual recovery criterion.



Figure 7: **Exhaustive joint semantic and perceptual recovery on PE generated states.** Exactly five of the 3,200 evaluated states, from five distinct classes, satisfy the fixed joint criterion, and all five are shown. Under both contracts, CTFM restores the matched-native ResNet-50 class missed by the calibrated base and contract-matched FFM-to-TLT while simultaneously improving PSNR, LPIPS, and semantic cycle error. The first column is the matched-native realization. Badges report PSNR and LPIPS to that realization. R followed by ✓/× indicates a correct/incorrect fixed ResNet-50 target-class prediction.

Fig. 8 extends the same strict recovery criterion to the two standard rendition shifts without refitting.

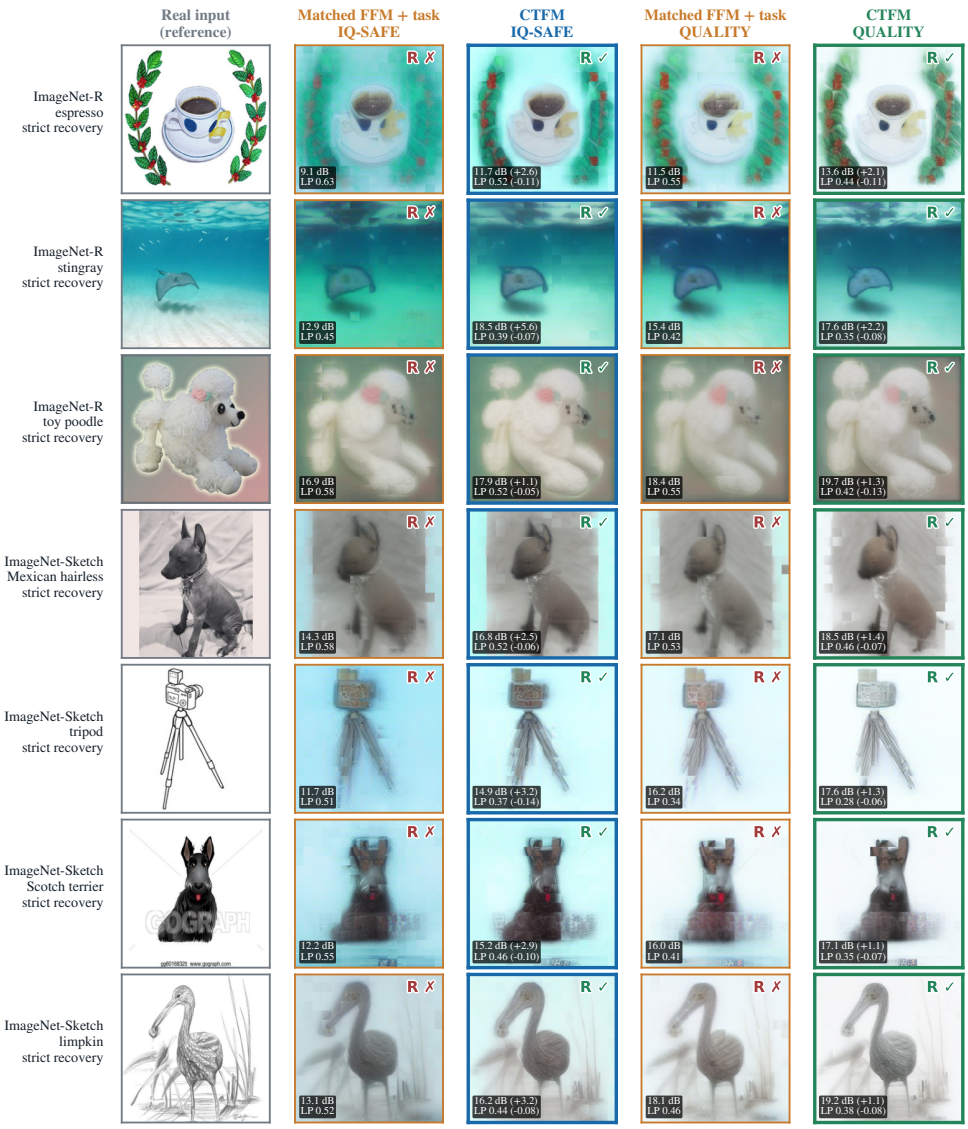


Figure 8: **Zero-refit joint semantic and perceptual recovery under standard rendition shifts.** Frozen PE maps are applied without refitting to 3,200 ImageNet-R and 4,000 ImageNet-Sketch images. Eligible examples require both CTFM contracts to restore the matched-native ResNet-50 decision missed by the contract-matched controls while improving PSNR, LPIPS, and semantic cycle error. We show all three eligible ImageNet-R classes and the four highest-scoring distinct classes among nine eligible ImageNet-Sketch classes; the minimax multi-metric ranking was fixed before rendering. Insets report PSNR and LPIPS to the real input, shown as the first-column reference. The remaining columns compare outputs of trained migration maps. R followed by ✓/× indicates a correct/incorrect fixed ResNet-50 target-class prediction. Population claims use all 7,200 images.

G Deployment and Reproducibility

G.1 Storage and Online Latency

Deployment accounting distinguishes weights trained in the current stage, all retained map weights, and complete serialized state, including fixed dependencies. MiB and GiB denote 2^{20} and 2^{30} bytes; MB denotes 10^6 bytes. These units are distinct from the accounting boundary (raw weights, stored tensors, or serialized files). Storage and runtime are evaluated separately. Tab. 21 combines complete serialized deployment state with end-to-end latency. Rank-432 CTFM and projected affine VFM

are matched to within 0.04% in storage, while CTFM is favorable in all six five-population macro endpoint families. Its complete online path adds 0.079 ms per state, or 1.04% over the shared archived decoder. Isolated-map operation counts and latency are retained in App. H.1.

Table 21: **RAEv2 deployment cost and performance.** System latency is map plus the archived decoder at batch eight on one RTX 4090; added latency is measured from the 7.576-ms decoder path. The family count reports favorable endpoint families averaged over five populations against projected affine VFM.

Map	Map weights	Total MB	Families	System ms	Added ms
Compact CTFM	163,840	5.518	–	7.609	+0.033
Rank-432 CTFM	1,048,576	**9.058**	**6/6**	7.654	+0.079
Projected affine VFM	1,049,600	9.061	0/6	**7.601**	**+0.025**

For rank-432 CTFM, decoding the identical precomputed field takes 7.576 ms per state. The complete online path therefore adds 0.079 ms, or 1.04% of decoder-only latency. The migration maps make no additional archived-consumer call before installed consumers execute the returned field, and they preserve the frozen generator, encoders, and decoder. The six-family macro improvement incurs this measured overhead; App. H.1 reports the corresponding isolated-map measurements.

Tab. 22 separates PE fitting time, trainable weights, stored tensors, and serialized checkpoint size.

Table 22: **PE deployment and fitting cost.** Both learned maps contain 163,840 trainable float32 weights. Raw weights, stored tensors, and serialized checkpoints use distinct accounting boundaries. The complete CTFM dependency state is 51.257 MiB, 30.94× smaller than retaining the 1.549-GiB matched decoder.

Quantity	CTFM	FFM-to-TLT
Refit time (s)	**38.00**	52.24
Selection plus refit (s)	**119.58**	159.15
Raw trainable-weight bytes	655,360	655,360
Stored tensor bytes	663,552	663,552
Serialized checkpoint bytes	671,241	671,241

The 655,360 raw weight bytes equal 0.625 MiB. Stored tensors add 8,192 bytes of fixed normalization buffers, and checkpoint serialization brings the file to 671,241 bytes, or 0.640 MiB.

The PE timing comparison uses recorded runtimes from runs on the same RTX 4090. It gives a 27.25% refit-time reduction and a 24.87% reduction for complete selection plus refitting, while the full-breadth CTFM map improves all six reported endpoint families. Timings are hardware-specific; parameter and byte counts are portable.

G.2 Reproduction materials

The supplement supplies per-sample results, model specifications, experimental partitions, evaluation instructions, and numerical consistency checks. The fitted map and its frozen dependencies determine deployment; evaluation models and checkpoint selection are used only offline.

H Sensitivity analyses and applicability

Sensitivity analysis examines computational accounting, class-level effects, optimization order, fitting breadth, response choice, and evaluation metrics under the declared deployment setting. The discussion distinguishes exact coordinate guarantees from empirical response improvements and specifies consumer types that require a different construction.

H.1 Isolated map-cost sensitivity

Tab. 23 separates the isolated field-map cost from the end-to-end deployed-system accounting in Tab. 21. The affine map is cheaper in isolation; the compiled CTFM path trades that marginal cost for

favorable effects in all six five-population endpoint families, while remaining a 1.04% addition to archived-decoder latency.

Table 23: **Isolated RAEv2 map cost.** Latency is the batch-16 median on one RTX 4090; learned storage excludes the shared frozen base.

Map	Learned MiB	MACs	Map ms
Compact CTFM	0.625	304.22M	0.040
Rank-432 CTFM	4.000	643.96M	0.073
Projected affine VFM	4.004	**269.55M**	**0.032**

H.2 Class-level recognition tail

Dataset-average recognition gains do not imply improvement in every class. The degraded-class proportions are 9.88%, 14.00%, 13.10%, 18.00%, and 15.50% on the source, ImageNet-R, ImageNet-Sketch, ImageNet-A, and ImageNetV2-D populations, respectively. In contrast, PSNR improves in at least 99.38% of classes and LPIPS in every class. Thus the degraded-class tail does not coincide with a general fidelity reversal; it reflects class-level movement around a frozen observer's decision boundary. The complete ECDFs in Fig. 3 expose every class and tail.

H.3 Paired optimizer-order robustness

Optimizer-order robustness is evaluated with three additional paired RAEv2 fits that hold the calibration population, rank-432 initialization, capacity, optimizer, projection, regularizer, checkpoint grid, and update budget fixed while changing the minibatch/optimizer seed. Each arm selects its own label-free validation minimum and is then evaluated on all 3,232 prospective states from 808 refinement-disjoint classes (App. B). Tab. 24 reports CTFM minus identical-capacity feature matching, with reductions for LPIPS and cycle MSE. The robustness evidence concerns optimization-order variation with fixed calibration data and upstream fitted dependencies.

Table 24: **Paired RAEv2 refinement robustness.** SD is the sample standard deviation across three independently ordered fits. Every effect is favorable in all three pairs.

Endpoint	Mean effect	SD	Minimum	Maximum	Fav. pairs
Gallery Top-1	+1.568 pp	0.143	+1.485	+1.733	3/3
ResNet-50 Top-1	+3.393 pp	0.064	+3.342	+3.465	3/3
EfficientNetV2-S Top-1	+2.712 pp	0.064	+2.661	+2.785	3/3
ViT-B/16 Top-1	+3.156 pp	0.031	+3.125	+3.187	3/3
PSNR	+1.858 dB	0.006	+1.853	+1.865	3/3
LPIPS reduction	0.05214	0.00004	0.05211	0.05218	3/3
Cycle-MSE reduction	0.12196	0.00011	0.12184	0.12205	3/3

H.4 Regional and response-component controls

Tab. 25 gives paired inference for the ownership-composition experiment in Tab. 13. Bootstrap draws resample the 808 classes; each contributes its first sample in the fixed dataset order. The compared maps are frozen and each control uses the same regional projector and retained anchor. All fourteen point effects favor CTFM. The seven sixteen-region intervals are strictly positive; the four-region Gallery interval contains zero.

Table 25: **CTFM versus matched feature refinement under regional ownership.** PSNR effects are in dB, recognition effects in percentage points, and LPIPS/cycle effects are reductions. Intervals use 20,000 paired bootstrap draws over the stated evaluation classes. Positive effects favor CTFM.

	Four regional means		Sixteen regional means	
Endpoint	**Effect**	**95% interval**	**Effect**	**95% interval**
Gallery Top-1	+1.114	[−0.866,+3.094]	+2.351	[+0.371,+4.332]
ResNet-50 Top-1	+5.322	[+3.094,+7.550]	+5.941	[+3.837,+8.168]
EfficientNetV2-S Top-1	+3.465	[+1.609,+5.322]	+4.208	[+2.351,+6.188]
ViT-B/16 Top-1	+3.342	[+1.114,+5.572]	+5.941	[+3.713,+8.168]
PSNR	+1.81091	[+1.70834,+1.91650]	+1.41447	[+1.34155,+1.48794]
LPIPS reduction	0.05615	[0.05336,0.05905]	0.05915	[0.05674,0.06162]
Cycle-MSE reduction	0.12787	[0.12194,0.13390]	0.13613	[0.13043,0.14210]

Tab. 26 complements the simultaneous-response analysis in Tab. 16. All fitted arms use the same 800-update horizon, seed, 640 fitting states, rank-432 factors, AdamW learning rate 10^{-5}, zero weight decay, batch four, and 10^{-4} displacement penalty. The normalization denominators use only initial errors on the fitting partition. The evaluation uses the first sample in fixed dataset order for each fit-disjoint class and no outcome-based checkpoint selection. RGB-only and cycle-only objectives use per-state denominators; the joint population control substitutes one fixed mean initial error for each response. The individual endpoints reveal the specialization of the single-response objectives, while simultaneous repair in Tab. 16 measures the joint objective's declared goal.

Table 26: **Individual endpoints of the response-component study.** Recognition is in percent, PSNR in dB; LPIPS and cycle are lower-is-better. All learned maps have 1,048,576 stored weights and use one fixed seed. Gray marks initialization, blue the joint per-state objective, and bold the best fitted value in each column.

Objective	**Gallery↑**	**ResNet↑**	**Eff.↑**	**ViT↑**	**PSNR↑**	**LPIPS↓**	**Cycle↓**
Initialization	52.23	73.51	77.10	75.25	16.884	0.2544	0.5946
Feature matching	52.85	73.14	75.74	76.11	16.244	0.2697	0.6152
RGB only	52.60	72.40	75.62	74.63	**18.017**	0.2763	0.6079
Cycle only	**53.34**	**75.62**	**78.71**	**76.98**	16.939	**0.2385**	**0.5409**
Joint, population normalization	52.85	73.76	77.72	76.24	17.562	0.2442	0.5611
Joint, per-state normalization	52.97	74.50	77.72	76.36	17.368	0.2431	0.5537

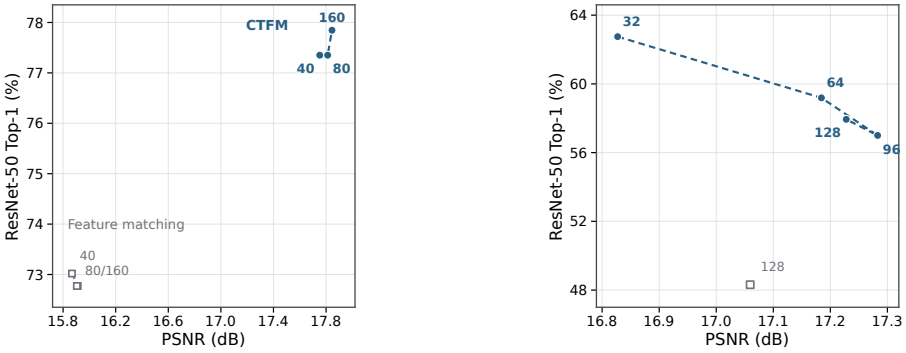
H.5 Residual/refinement-fit breadth sensitivity

Tab. 27 reports sensitivity to the number of fitting classes, retaining the complete breadth sweeps behind the data-efficiency analysis. Each host uses one fixed evaluation population and identical deployment capacity within its comparison. The cross-breadth claims in Tab. 14 use separately frozen prospective comparisons. The trajectories document data sensitivity under each method's prescribed checkpoint-selection rule.

Table 27: **Residual/refinement-fit breadth sensitivity.** The frozen release transport is shared across rows. Eff./ViT are independent Top-1 accuracies. Lower LPIPS and cycle MSE are better.

Host	Method / classes	Gallery	ResNet-50	Eff.	ViT	PSNR	LPIPS	Cycle
RAEv2	Feature matching / 40	53.280	73.020	76.856	75.619	15.867	0.29337	0.65583
	Feature matching / 80	53.589	72.772	77.228	75.557	15.911	0.29220	0.65379
	Feature matching / 160	53.775	72.772	77.104	75.557	15.904	0.29195	0.65230
	CTFM / 40	55.012	77.351	79.332	78.713	17.752	0.24466	0.54840
	CTFM / 80	55.322	77.351	79.703	79.022	17.812	0.24056	0.53853
	CTFM / 160	**55.569**	**77.847**	**79.827**	**79.579**	**17.845**	**0.23661**	**0.53056**
PE	FFM-to-TLT / 128	25.875	48.313	53.875	52.438	17.059	0.49537	18.74895
	CTFM / 32	29.188	62.750	66.312	65.375	16.827	0.38283	18.27612
	CTFM / 64	27.937	59.188	63.563	62.437	17.184	0.40479	18.43505
	CTFM / 96	27.062	57.000	62.187	59.312	17.283	0.41460	18.48703
	CTFM / 128	28.313	57.938	62.250	61.187	17.228	0.42303	18.44016

Fig. 9 visualizes the recognition–fidelity coordinates from Tab. 27 immediately below. Each point is a separately fitted map with its own training trajectory.



(a) RAEv2 response refinement.

(b) PE residual fitting.

Figure 9: **Residual-stage data controls the recognition–fidelity trade-off.** Circles denote CTFM, squares denote the control, and labels give fitting classes. Dashed segments connect observed fits in increasing-budget order and do not imply interpolation; PE has only the measured 128-class control. Upper-right is better within each host and axis ranges differ.

H.6 Distribution-estimator sensitivity

Tab. 28 checks the distribution comparison with KID, whose feature-distribution estimator differs from FID as defined in App. D.5. Reduced-data CTFM improves native KID on both hosts and real-reference KID on PE. On RAEv2, real-reference KID differs by 1.37×10^{-5} in the opposite direction, while real-reference FID and all reported classwise real-reference feature distances improve.

Table 28: **Secondary KID sensitivity.** Lower is better; full and reduced refer to the residual-stage data fractions defined in Tab. 15.

Host	Method	Native KID	Real KID
RAEv2	Feature matching / full	0.006449	**0.0014480**
	CTFM / reduced	**0.003576**	0.0014617
PE	FFM-to-TLT / full	0.057788	0.032944
	CTFM / reduced	**0.047409**	**0.024623**

Tab. 29 gives uncertainty for the classwise feature-distance reductions on both hosts. These quantities are computed per class and then averaged, which allows resampling classes for inference.

Table 29: **Class-decomposable distribution-risk reductions.** Positive values favor limited-residual-data CTFM over the full-data matched-capacity baseline. Intervals use 20,000 class-cluster bootstrap draws.

Host	Endpoint	Reduction	95% interval
RAEv2	Paired Inception MSE	0.009940	[0.009258,0.010646]
	Native class-centroid MSE	0.004071	[0.003742,0.004421]
	Symmetric real adequacy	0.003765	[0.003305,0.004240]
	Real class-centroid MSE	0.001051	[0.000756,0.001349]
PE	Paired Inception MSE	0.017737	[0.015916,0.019591]
	Native class-centroid MSE	0.011958	[0.010599,0.013344]
	Symmetric real adequacy	0.010295	[0.008869,0.011734]
	Real class-centroid MSE	0.006476	[0.005390,0.007558]

H.7 Frozen vision–language observer

A frozen SigLIP2-B/16 evaluation model (Tschannen et al., 2025) measures vision–language transfer through image similarity to class-name text embeddings. This gallery-free evaluation uses all 10,000 MatchedFrequency images for ImageNetV2. The model is excluded from fitting and selection. Against compact FFM-to-TLT and identical-capacity feature matching, CTFM improves all six domain-macro vision–language endpoints across the source and four ImageNet shifts; zero-shot Top-1/Top-5 gains are $+11.10/+15.36$ and $+4.03/+3.86$ points, respectively. Against the stronger projected affine VFM map, Tab. 30 shows strict gains on five endpoints; the true-class text-cosine interval spans zero, leaving that difference unresolved.

Table 30: **Frozen SigLIP2 observer against projected affine VFM.** Effects are CTFM minus VFM under five-population equal weighting. Prediction metrics are percentage-point changes; intervals use 20,000 class-within-domain bootstrap draws.

Endpoint	Effect	95% interval
Zero-shot Top-1	+0.532 pp	[+0.107,+0.963]
Zero-shot Top-5	+0.529 pp	[+0.163,+0.892]
Ground-truth text margin	+0.000650	[+0.000504,+0.000800]
Native-image cosine	+0.005270	[+0.004906,+0.005630]
Native-prediction agreement	+1.378 pp	[+0.887,+1.873]
True-class text cosine	−0.000069	[−0.000208,+0.000071]

H.8 Actuator topology and optimization horizon

The compact actuator combines global token exchange with paired initialization. Tab. 31 tests these two design choices on a fixed 128-class PE development population. At 300 steps, global token exchange outperforms local depthwise alternatives, and paired cancellation improves all five endpoints over a standard zero-output global actuator at identical parameter count. The longer-horizon rows show that initialization also changes the attainable trade-off, motivating checkpoint selection using the complete declared functional.

Table 31: **Matched PE actuator study.** All learned rows use the same IQ-SAFE functional. The 300-step block varies topology and initialization; the 900-step block tests horizon at matched global capacity. Blue marks the released paired initialization.

Actuator	Steps	Rank / params	Gallery ↑	ResNet-50 ↑	PSNR ↑	LPIPS ↓	Cycle ↓
300 updates: topology and initialization							
No learned correction	0	– / 0	22.27	41.80	14.531	0.4507	19.830
Local depthwise	300	32 / 65.8K	21.88	41.02	14.493	0.4499	19.799
Local depthwise	300	128 / 263.3K	21.68	41.80	14.524	0.4499	19.796
Global zero-output	300	32 / 163.8K	22.46	43.36	14.614	0.4450	19.659
Global paired	300	2×16 / 163.8K	**23.44**	**46.29**	**14.620**	**0.4404**	**19.345**
900 updates: matched global capacity							
Global zero-output	900	32 / 163.8K	25.98	**53.13**	**15.025**	**0.4206**	18.623
Global paired	900	2×16 / 163.8K	**27.93**	50.98	14.619	0.4354	**18.490**

At 300 steps, paired cancellation improves Gallery by 0.98 points, ResNet-50 by 2.93 points, PSNR by 0.006 dB, LPIPS by 0.00465, and cycle MSE by 0.3138. The result supports the specific design claim: paired branches have an exact zero-correction initialization with nonzero first-order gradients.

H.9 TANGENT INITIALIZATION ACROSS RANKS

The rank controls the number of channel directions retained by the compiled correction. Tab. 32 reports the rank profile from the fixed development grid in App. A.5. The coefficient is held at $\alpha = 1$ for this comparison; the selection itself considered all 72 rank–coefficient pairs, of which 54 were admissible. The selected rank balances the two semantic endpoints before perceptual and pixel fidelity, following the prescribed multi-metric order. These are initialization results, before response-guided factor refinement.

Table 32: **Development rank profile of the ownership-feasible initialization.** All rows use the same 128 development states and $\alpha = 1$. Weights include the retained anchor. Blue marks the selected initialization; recognition is in percent and PSNR in dB.

Rank	Weights	Gallery↑	ResNet-50↑	PSNR↑	LPIPS↓	Cycle↓
32	229,376	44.531	58.594	15.694	0.34665	0.86972
64	294,912	51.562	71.875	16.064	0.30994	0.75039
128	425,984	52.344	75.000	16.418	0.28377	0.67175
192	557,056	54.688	71.875	16.582	0.27145	0.64432
224	622,592	55.469	71.875	16.608	0.26764	0.63471
256	688,128	55.469	72.656	16.680	0.26314	0.62824
320	819,200	54.688	71.875	16.801	0.25773	0.61739
384	950,272	53.906	73.438	16.866	0.25474	0.60844
432	1,048,576	56.250	74.219	16.902	0.25134	0.60132

H.10 OWNERSHIP INTERVENTION

The matched PE intervention isolates the ownership constraint by changing only whether the learned correction is projected. Removing projection improves ResNet-50 by 3.81 points and every decoded-image endpoint, while pooled RMS drift rises from 4.77×10^{-8} to 1.582. With projection, the learned correction contributes only 2.87×10^{-9} pooled RMS beyond the calibrated base and changes no pooled Gallery decision. This controlled intervention supplies the empirical counterpart of Eq. (6): quotient invariance is exact, and its quality cost is measurable.

Tab. 33 reports the corresponding ownership-free frontier and pooled-coordinate drift.

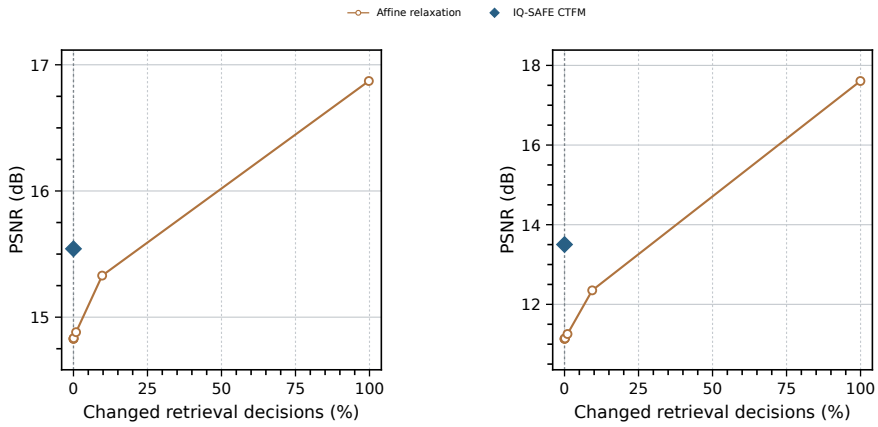
Table 33: **Ownership-free PE sensitivity.** These maps operate without the pooled-coordinate guarantee. Blue marks CTFM and orange marks the full affine alignment; maximum drift is measured from the base-calibrated pooled coordinate.

Method	Gallery ↑	ResNet-50 ↑	PSNR ↑	LPIPS ↓	Cycle ↓	Max drift ↓
FFM-to-TLT	29.72	52.00	17.278	0.4515	17.962	6.581
QUALITY CTFM	30.47	58.25	**17.333**	0.3890	17.736	9.555
Affine VFM	**47.63**	**78.72**	16.176	**0.2861**	**15.161**	67.630

The PE residual gap is not explained solely by ownership: removing projection raises ResNet-50 from 54.44 to 58.25, while the archived-field reference is 84.13. The ownership-free affine row also changes parameterization and fitting conditions, so these results do not isolate whether capacity, target choice, or release mismatch dominates the remaining gap. Deployment therefore retains a map only when its installed consumers satisfy release-specific validation criteria.

Continuous ownership relaxation. Continuous ownership relaxation interpolates the pooled output of a fixed affine map while preserving its centered spatial output. On the same 3,200 ImageNet-R and 4,000 ImageNet-Sketch queries as Tab. 17, we hold the affine map V fixed and vary only the fraction of its pooled displacement that is allowed. With $q_0 = Q(A_0(Z))$, the returned field is $A_t(Z) = P_0V(Z) + \mathbf{1}[(1-t)q_0 + tQ(V(Z))]$ for the fixed grid $t \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The endpoints are projected and unrestricted affine VFM. Each setting is release-wide; no weights, checkpoints, or settings are selected using these queries. The original IQ-SAFE CTFM map is a separate fixed comparator. All arms use the same clamped query tensor and direct-RGB evaluation.

Fig. 10 quantifies the operational effect of coordinate ownership. Releasing the full affine mean displacement changes 99.78% and 99.95% of installed top-ranked decisions, respectively. CTFM changes none and improves PSNR over the projected affine endpoint by 0.71 and 2.37 dB; its LPIPS is also lower in both domains. Intermediate settings expose a graded trade-off between coordinate preservation and image fidelity.



(a) ImageNet-R. (b) ImageNet-Sketch.

Figure 10: **Image fidelity versus installed retrieval changes.** Circles join the six evaluated relaxations of one frozen affine map; the diamond is the original IQ-SAFE CTFM map. The horizontal axis counts changes from the calibrated base's top-ranked item. These are decision-disagreement counts. Higher PSNR and fewer changed decisions are preferred. Near-zero settings can overlap because their effects are small; no setting is omitted.

H.11 FINITE-PRECISION RETRIEVAL STABILITY

Exact mean preservation holds algebraically. Under floating-point execution, a sufficient condition for retrieval stability bounds score perturbations relative to the decision margin. The analysis compares the gap between each query's largest and second-largest scores with a perturbation bound.

For a fixed gallery with vectors c_j, let u and u' be the normalized calibrated-base and migrated queries. Their exact score differences satisfy $|\langle u' - u, c_j \rangle| \le G\|u' - u\|_2$, where $G = \max_j \|c_j\|_2$.

Let γ be the computed baseline winner–runner-up score gap, and let η_0, η_1 bound the per-score arithmetic errors for the two queries. A sufficient condition for an unchanged top-ranked item is

$$\gamma > 2\big(G\|u' - u\|_2 + \eta_0 + \eta_1\big). \tag{28}$$

The result follows by bounding the change of each of the two competing scores. The margin compares the two highest-scoring gallery items independently of the ground-truth class. Gallery size can affect how small this margin becomes; independent field-map applications do not themselves accumulate error across queries. Float64 reference normalization and dot products provide a numerical check on the supported float32 scorer, while worst-case arithmetic bounds and approximate search behavior require their own implementation analysis.

Tab. 34 evaluates this condition on all 7,200 query/gallery-disjoint images. Both the normalized query perturbation and the float32 score discrepancy from float64 reference computation are measured per query. Every query satisfies Eq. (28); no top-ranked item changes relative to the calibrated base, and float32 versus float64 scoring also changes zero decisions. Maximum pooled-coordinate error remains 4.77×10^{-7} in both domains. These measurements check the declared 1,000-centroid scorer; the sufficient condition makes the role of score margins explicit without assuming a particular future index size.

Table 34: **Empirical retrieval-margin check under float32 execution.** The bound is the right-hand side of Eq. (28), using measured query perturbations and float64-reference score discrepancies. Margins and bounds are in cosine-score units. Passes count queries satisfying the inequality; flips count changed top-ranked items relative to the base.

Population	Queries	Min. margin	Max. bound	Passes	Flips
ImageNet-R	3,200	1.36×10^{-5}	3.33×10^{-7}	3,200	0
ImageNet-Sketch	4,000	3.83×10^{-6}	2.92×10^{-7}	4,000	0

H.12 Applicability of the declared contract

The declared contract determines the guarantee's scope. Owned-coordinate invariance, response-error control, and discrete-decision stability are distinct properties.

Choosing behavior and ownership. The functional F and statistic O express separate deployment requirements and can serve different consumers. For decoder repair with a retained gallery, F measures decoder-side responses and $O = Q$ preserves the calibrated query. Choosing $O = I$ instead protects every field coordinate: $\ker I = \{0\}$ then forces the added correction to vanish. An ownership rule therefore specifies which degrees of freedom remain available for repair.

Task heads and evaluation models. Every deterministic head $\psi \circ Q$, including a linear or MLP classifier, receives exactly the same pooled input as under A_0 in the IQ-SAFE contract. A head $r \circ F$ instead inherits the response-error bound when r satisfies the Lipschitz condition in App. A.3; discrete predictions also depend on the decision margin. The experiments exercise archived decoding, pooled retrieval, and regional descriptors. ResNet-50, EfficientNetV2-S, ViT-B/16, and the vision–language observer evaluate decoded images; these measurements do not constitute deployment tests of task heads on migrated fields.

Conditioned consumers. ControlNet introduces spatial conditioning into a diffusion model (Zhang et al., 2023). Applying a consumer-typed contract to such a system would require specifying the consumer's compatible input field, conditioning signals, and diffusion timesteps when defining F. Exact coordinate protection would also require a suitable linear statistic O and its nullspace projector. This is a potential extension of the formulation, separate from the RAE consumer evaluations reported here.

The local surrogate bound in Eq. (11) is conditional on its stated smoothness assumptions. The interventions do not estimate β or ϵ; discrete Top-1 readouts and the globally nonsmooth maximum in Eq. (9) are not assigned that smooth-risk guarantee.

CTFM addresses post-release migration between same-width dense fields. A release declares the archived functional and any owned coordinates before held-out evaluation; deployment then applies

Table 35: **RAEv2 gains after excluding compiler-selection classes.** Both comparisons use 3,104 states from 776 classes. Recognition effects are percentage points; LPIPS and cycle entries are reductions. Positive effects favor CTFM. Intervals use 20,000 paired class-cluster bootstrap draws.

	Equal refinement data		Quarter refinement data	
Endpoint	**Gain**	**95% interval**	**Gain**	**95% interval**
Gallery Top-1	**+1.740**	[+0.773,+2.706]	**+1.546**	[+0.612,+2.481]
ResNet-50 Top-1	**+3.640**	[+2.577,+4.704]	**+3.963**	[+2.899,+5.026]
EfficientNetV2-S Top-1	**+2.642**	[+1.611,+3.673]	**+2.610**	[+1.643,+3.608]
ViT-B/16 Top-1	**+3.415**	[+2.223,+4.607]	**+2.738**	[+1.740,+3.769]
PSNR (dB)	**+1.862**	[+1.799,+1.928]	**+1.821**	[+1.761,+1.882]
LPIPS reduction	**0.05234**	[0.05056,0.05415]	**0.04568**	[0.04401,0.04738]
Cycle-MSE reduction	**0.12257**	[0.11912,0.12610]	**0.10450**	[0.10107,0.10800]

one map shared across inputs without per-input optimization. The functional-coverage result in App. A.3 states precisely which future consumers inherit the response guarantee. The construction accepts any linear ownership operator with a computable nullspace projector. The experiments validate global token-mean ownership and composition with four or sixteen regional means (App. D.3). Attention pooling, max pooling, and other nonlinear ownership rules require a distinct projector or an approximate contract. Width-changing releases and consumers not expressible through the declared functional require a new contract and a new held-out admission test. Specifically, if $F(Z) = F(Z')$ for the declared functional but an undeclared consumer satisfies $u(Z) \neq u(Z')$, no vanishing global bound for that consumer can follow through d_F because its right-hand side is zero. For a discrete decision derived from a covered response, unchanged Top-1 additionally requires its response bound to remain below the consumer's decision margin.

The reported pipeline uses E_{old} and D_{new} offline, or requires cached equivalent paired targets and base dependencies. A decoder-only or unpaired configuration is not established by these experiments. Because the response targets derive from matched-native realizations, the migrated field can inherit their representational and image-generation biases; real-reference distribution metrics and observers excluded from fitting provide external checks; inherited artifacts may persist despite favorable evaluation metrics.

H.13 Compiler-selection sensitivity

The fresh-sample evaluations and upstream selection history are specified in App. B. We remove the 32 compiler-selection classes from each prospective RAEv2 comparison, leaving 776 classes and 3,104 states. These populations exclude calibration and compiler-selection classes but retain the 128 earlier objective-selection classes. This post-hoc analysis therefore isolates compiler-selection overlap. Exclusion uses class provenance alone; no checkpoint, threshold, or example is selected using the resulting scores, and all maps remain frozen.

All seven point estimates and their intervals remain positive in both comparisons (Tab. 35). The original full-population results are retained with their stated selection scope; the complementary analysis confirms that their direction persists after compiler-selection exclusion.