RST: ROUTES FOR CERTIFYING AND CONSTRUCTING PRESCRIBED INFORMATION IN VARIATIONAL AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Posterior-collapse diagnostics can establish that a VAE uses its latent representation, but they do not determine whether a specified information view crosses a representation/readout interface fixed before training. Reconstruction–Student–Teacher (RST) makes that stronger requirement explicit: **R denotes Reconstruction**, S the fixed Student witness, and T the frozen Teacher specification. Let $T_X = T(\cdot \mid X)$ be a full-support categorical Teacher, $C \mid X \sim T_X$, $\bar{T} = \mathbb{E}_X T_X$, and let an audited representation Z satisfy $C - X - Z$. With $I_T := I(C; X)$, $L_{\mathrm{TS}} := \mathbb{E}\,\mathrm{KL}(T_X \| S(\cdot \mid Z))$, and $G_T := I_T - L_{\mathrm{TS}}$, we prove $I(C; Z) \geq G_T$ and $L_{\mathrm{TS}} \geq I_T$ for every input-independent $Z \perp X$, with equality attained by the centered RST witness. Thus $G_T > 0$ certifies transmission of the declared Teacher information through the specified interface. In the main experiments $Z = u_\phi(X)$ is the deterministic encoder mean, while reconstruction uses $z \sim q_\phi(z \mid x)$.

A centered regular-simplex witness turns the certificate into an explicit construction: a minimum-dimensional closed-form Teacher code, its complete affine solution fiber, exact witness-visible/witness-null routing, a margin–energy path, strict function-space departure from the centered input-independent point, certified preservation cylinders, and orthogonal multi-view composition. Five-dataset experiments test boundary crossing, stronger KL pressure, Teacher counterfactuals, witness-null routing, escape, and preservation. On CIFAR-100, learned-head auxiliary (Aux) codes are probe-readable while their predeclared margins are negative. Cross-seed Aux-head reuse falls from $36.57/21.88\%$ native accuracy to $5.20/0.94\%$, whereas one fixed RST witness directly reads all five Conv-RST encoders at $43.84/33.58\%$. The same witness is also certificate-positive on five independently optimized ResVAE RST encoders, with coarse/fine margins $G_c = 0.869 \pm 0.057$ and $G_f = 1.083 \pm 0.105$; Conv↔Res learned-head reuse is near chance, while a ridge-affine adapter makes all $100/100$ transferred margins positive. These tests validate a functional consequence of prescribed alignment; they do not claim fixed simplex classifiers or compatible representations as new.

1 INTRODUCTION

A variational autoencoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014) can fail to use its representation in more than one way. The posterior may become input-independent; the code may remain active while discarding a factor of interest; or the factor may be present but only in an arbitrary orientation that must be discovered by fitting a new probe. These cases are not equivalent. Ordinary downstream prediction often needs only the last property, but modular or auditable representations can require more: a route and its readout geometry may have to be declared before training and then reused, checked, or composed without learning a new coordinate system. We call this requirement *prescribed alignment*.

Intuition. The frozen Teacher says *what* should be readable; the fixed Student witness says *how* it must be read; and the encoder must make that predeclared reader succeed. The certificate asks whether this success is better than any input-independent code. Figure 3 then turns the same test into geometry: visible directions carry the declared view, witness-null directions leave residual freedom, and an analytic path moves from the centered reference toward an exact Teacher code.

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.

Table 1: What the exact input-independent boundary leaves open, and what the simplex construction adds.

Question after certification	Construction developed here
How can the declared Teacher be realized?	Closed-form minimum-norm Teacher code (denoted $u_T(x)$ below) in the minimal $K-1$ witness-visible dimensions
Where can residual variation live?	Complete affine solution fiber and exact certificate invariance along witness-null directions
How can one leave the input-independent reference, and at what cost?	Analytic interpolation to the Teacher code, with a certified margin, quadratic mean-prior energy, and a strict local departure direction
How can an informative state survive later training?	A visible-space preservation radius with unrestricted witness-null motion
How can several declared views coexist?	Orthogonal simplex blocks with simultaneous margins and additive visible-energy budgets

Formally, RST names three functional relations: **R=Reconstruction**, the ordinary VAE reconstruction path; **S=Student**, the fixed witness/readout geometry; and **T=Teacher**, the frozen information specification. Throughout, Z denotes the audited representation (main experiments: $Z = Z_\mu := u_\phi(X)$), while lowercase $z \sim q_\phi(z \mid x)$ is the sampled latent sent to the decoder. With $T_X := T(\cdot \mid X)$, $C \mid X \sim T_X$, $\bar{T} := \mathbb{E}_X T_X$, and $C - X - Z$, define

$$I_T = \mathbb{E}_X \mathrm{KL}(T_X \| \bar{T}), \qquad L_{\mathrm{TS}} = \mathbb{E}_{X,Z} \mathrm{KL}(T_X \| S(\cdot \mid Z)), \qquad G_T = I_T - L_{\mathrm{TS}}.$$

Here I_T measures declared Teacher information, L_{TS} fixed-reader mismatch, and G_T the certificate margin; soft training uses $\mathcal{L}_{\mathrm{rec}} + \beta_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{TS}} L_{\mathrm{TS}}$ with frozen Teacher and witness. The exact decomposition below gives $I(C;Z) \geq G_T$. Every input-independent representation $Z \perp X$ satisfies $L_{\mathrm{TS}} \geq I_T$, and the centered simplex witness attains equality at the constant origin. The corresponding best-constant/input-independent boundary was established in the earlier Teacher-guided certificate (Zhang et al., 2026); here it is the reference point for the constructive geometry, not a new boundary claim. Input-independent means $Z \perp X$, distinct from the later witness-null directions, which are invisible to the fixed witness but need not be input-independent. Seen through this formulation, exact constant posterior collapse is the canonical extreme case: an input-independent representation cannot transmit any nontrivial declared Teacher view. Hence $G_T > 0$ rules out that collapse on the audited route, while the broader prescribed-information formulation also distinguishes failures that ordinary non-collapse diagnostics can miss. Posterior collapse is therefore the central input-independent reference case, not the full scope of the theory.

The central question is what becomes possible once this input-independent reference boundary is fixed. A boundary can tell us that a representation has crossed input independence, but it does not tell us how to realize a chosen Teacher, how much visible capacity that realization needs, what degrees of freedom remain available, how to depart from the input-independent reference and quantify the cost, whether that centered reference has a strict departure direction, how a positive-margin state can survive later optimization, or how several independently specified views can coexist. For a full-support Teacher with K classes, the regular-simplex construction answers these questions in one geometry:

$$\boxed{\text{CERTIFY} \rightarrow \text{REALIZE} \rightarrow \text{ROUTE} \rightarrow \text{MOVE/COST} \rightarrow \text{ESCAPE} \rightarrow \text{PRESERVE} \rightarrow \text{COMPOSE}}$$

Table 1 is the roadmap for Section 2. The ingredients—KL decompositions, softmax log-odds, regular simplices, tight frames, and exponential interpolation—are classical; the contribution is the way they are assembled into a fixed prescribed-information interface with exact realizations, auditable witness-null directions, explicit energy budgets, and composable protected routes. Under centered unit-row minimal-dimensional linear-softmax witnesses, requiring minimum visible dimension, an explicit certificate-invariant residual complement, and minimum normalized worst-case sensitivity selects the regular simplex uniquely up to rotation (Appendix B.3); the remaining construction then operates on this same geometry.

The empirical program mirrors this dependency rather than optimizing a single benchmark score. Five image datasets test boundary crossing; KL stress, Teacher counterfactuals, witness-null routing, escape, and certified refinement probe distinct parts of the theory. CIFAR-100 then separates

$$\boxed{\text{latent activity} \neq \text{learned decodability} \neq \text{prescribed alignment.}}$$

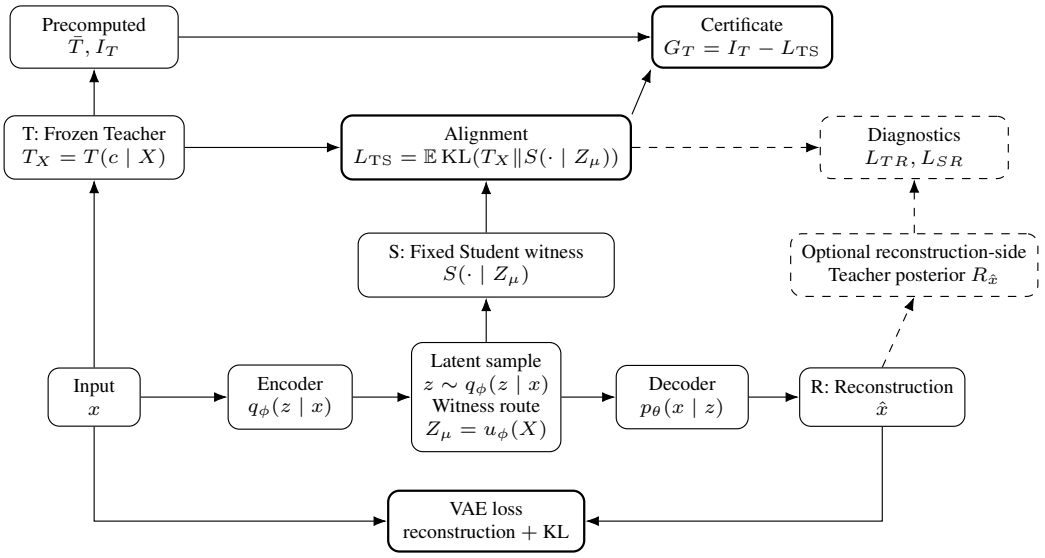

Figure 1: RST architecture (R=Reconstruction, S=Student, T=Teacher). The decoder sample z is distinct from the audited mean $Z_\mu = u_\phi(X)$. Here $R_{\hat{x}}$ is the frozen Teacher posterior on the reconstruction; L_{TR} and L_{SR} compare Teacher/Student with it. These dashed diagnostics are defined in Appendix E.4.5 and do not enter the certificate.

The separation is both formal and functional. Appendix C.1 gives an exact rotation counterexample. In matched controls, frozen-encoder probes recover the Teacher while the predeclared Aux interfaces remain certificate-negative; learned Aux heads fail almost completely under both cross-seed and Conv↔Res no-refit reuse, but small adapters recover much of the lost performance. In contrast, the same predeclared RST witness directly reads all five ConvVAE seeds and a second independently optimized ResVAE backbone. RST therefore targets a narrower no-refit interface contract, not ordinary decodability or compatibility as a field-level novelty. Figure 1 shows these three routes.

Main contributions.

1. **A route-local certificate written directly on the representation variable.** For a frozen Teacher and a witness strictly positive on its support we give the exact information decomposition, the attainable input-independent boundary for the centered witness, and a re-instantiation principle for any chosen Z_m satisfying the corresponding Markov condition. Here Z_m is a mathematical representation variable and is distinct from the R in RST, which denotes Reconstruction.
2. **A canonical prescribed-interface geometry.** For a full-support K-class Teacher we derive the minimal $K-1$ witness-visible dimension, the unique minimum-norm Teacher code, the complete affine solution fiber, and exact visible/witness-null routing. Under the stated normalized minimal-interface criterion, the simplex is unique up to rotation; simplices themselves are not claimed as new.
3. **Movement, cost, escape, and preservation in the same geometry.** The analytic Teacher ray has a certified margin–energy law and a strictly negative alignment-loss derivative at the centered input-independent point, while a Lipschitz bound gives a positive-margin preservation cylinder around informative anchors.
4. **Composable prescribed views with functional interface controls.** Orthogonal simplex blocks provide simultaneous route-wise guarantees and additive energy budgets. CIFAR-100 recovers the predicted selective sign pattern in all five seeds; frozen probes show information can survive when the declared interface fails, and cross-seed/cross-architecture reader transfer localizes the failure to producer–consumer coordinates. The same predeclared witness is reused without refitting across ConvVAE and ResVAE RST encoders; small adapters and controlled gauge/zero/swap interventions isolate the corresponding interface behavior.

The claim boundary is intentionally explicit. A positive G_T is a sufficient certificate relative to the declared Teacher, route, witness, and evaluation distribution. It does not prove semantic completeness, decoder use of that route, disentanglement, parameter-space convergence, or the absence of every other form of posterior collapse.

2 FROM THE EXACT BOUNDARY TO CONSTRUCTIVE SIMPLEX GEOMETRY

Exact route-local certificate. In plain terms, this asks whether the fixed reader can beat every input-independent representation.

Theorem 1 (Information decomposition and attainable input-independent boundary)**.** *For any $C - X - Z$ and witness $S(C \mid Z)$ that is strictly positive on the Teacher support,*

$$\begin{aligned} L_{\mathrm{TS}} &= I(C; X \mid Z) + \mathbb{E}_Z \mathrm{KL}(P(C \mid Z) \| S(C \mid Z)), \\ G_T &= I(C; Z) - \mathbb{E}_Z \mathrm{KL}(P(C \mid Z) \| S(C \mid Z)). \end{aligned} \tag{1}$$

Hence $I(C; Z) \geq G_T$. If $Z \perp X$, then $P(C \mid Z) = \bar{T}$ and

$$L_{\mathrm{TS}} = I_T + \mathbb{E}_Z \mathrm{KL}(\bar{T} \| S(C \mid Z)) \geq I_T.$$

If the witness admits a constant code u_0 with $S(\cdot \mid u_0) = \bar{T}$—as the centered simplex witness below does at $u_0 = 0$—the boundary is attained. Thus, for the RST witness,

$$\boxed{G_T > 0 \Longrightarrow I(C; Z) \geq G_T > 0, \qquad \inf_{Z \perp X} L_{\mathrm{TS}} = I_T.}$$

A nonpositive margin is inconclusive.

Corollary 1 (Route-local re-instantiation)**.** *Replacing (C, Z, T, S) by $(C_m, Z_m, T^{(m)}, S_m)$ gives the same exact boundary and sufficient information margin for any chosen representation/statistic satisfying $C_m - X - Z_m$, provided the witness has a constant code producing the corresponding marginal.*

The identities also hold exactly on a fixed finite split when $\bar{T}$, I_T, and L_{TS} are computed self-consistently on that empirical distribution; this is not a population-confidence statement. Proofs and the population extension are in the appendix.

Canonical interface geometry: realization, freedom, and stability. The next question is geometric: what is the shortest exact Teacher code, and which directions can vary without changing the fixed readout? Requiring minimum visible dimension, an explicit witness-null complement, and minimum normalized worst-case sensitivity uniquely selects regular-simplex rows up to rotation among centered unit-row minimal-dimensional witnesses (Appendix B.3). For a realized input x, write $T_x := T(\cdot \mid x)$. Let $v_1, \ldots, v_K$ be unit regular-simplex vertices, $\sum_k v_k = 0$, $v_i^\top v_j = -1/(K-1)$ for $i \neq j$, and let V collect their rows. With gain $\beta > 0$ define

$$S(u) = \mathrm{softmax}(\beta V u + \log \bar{T}), \qquad \mathcal{V} = \mathrm{row}(V), \quad \mathcal{N}_V = \ker V = \mathcal{V}^\perp.$$

Let $P_{\mathcal{V}}$ denote the orthogonal projector onto $\mathcal{V}$.

Proposition 1 (Canonical simplex interface: minimality, exact fiber, and witness-null invariance)**.** *A linear-softmax witness that locally represents all full-support K-class categorical distributions needs at least $K - 1$ visible dimensions; the simplex attains this minimum. For full-support T_x, define the centered log ratio*

$$r_k(x) = \log \frac{T_{x,k}}{\bar{T}_k} - \frac{1}{K} \sum_j \log \frac{T_{x,j}}{\bar{T}_j}, \qquad u_T(x) = \frac{K-1}{K\beta} V^\top r(x).$$

Then

$$\boxed{S(u_T(x)) = T_x, \qquad S^{-1}(T_x) = u_T(x) + \ker V,} \tag{2}$$

$u_T(x)$ is the unique minimum-norm realization, and

$$\frac{1}{2} \mathbb{E} \| u_T(X) \|^2 = \frac{K-1}{2K\beta^2} \mathcal{E}_T, \quad \mathcal{E}_T = \mathbb{E} \| r(X) \|^2.$$

Writing $u_{\mathcal{V}} := P_{\mathcal{V}} u$ and $u_{\mathcal{N}_V} := u - u_{\mathcal{V}}$, every $u = u_{\mathcal{V}} + u_{\mathcal{N}_V}$ satisfies $S(u) = S(u_{\mathcal{V}})$: witness-null variation is exactly certificate invariant, although it still pays ordinary VAE KL cost and need not be used by the decoder. The regular simplex is not required by the certificate theorem; it is a canonical choice because it simultaneously gives minimal visible dimension, class symmetry, an isotropic tight frame, the closed-form inverse/fiber, an orthogonal witness-null complement, and the smallest normalized worst-case sensitivity among unit-norm row-centered minimal-dimensional witnesses.

Movement, cost, and function-space departure. Next we ask how to leave the centered reference, and at what Gaussian mean-prior energy cost.

Proposition 2 (Analytic margin–energy path)**.** *Along* $u_\alpha(x) = \alpha u_T(x)$, $0 \le \alpha \le 1$, *define* $G(\alpha) := I_T - L_{\mathrm{TS}}(u_\alpha)$ *and* $E_{\mathrm{mean}}(\alpha) := \frac{1}{2}\mathbb{E}\|u_\alpha(X)\|^2$. *Then*

$$S_k(u_\alpha) = \frac{\bar{T}_k^{1-\alpha} T_{x,k}^{\alpha}}{\sum_j \bar{T}_j^{1-\alpha} T_{x,j}^{\alpha}}, \qquad \boxed{G(\alpha) \ge \alpha I_T}, \qquad \boxed{E_{\mathrm{mean}}(\alpha) = \alpha^2 E_{\mathrm{mean}}(1)}. \tag{3}$$

G is concave and nondecreasing, with $G(0) = 0$, $G(1) = I_T$, and initial slope $\mathbb{E}[J(T_X, \bar{T})]$, where $J(P, Q) := \mathrm{KL}(P\|Q) + \mathrm{KL}(Q\|P)$ is the Jeffreys divergence.

Corollary 2 (Strict function-space escape from the centered input-independent solution)**.** *Assume $I_T > 0$ and full support. Within the class of input-independent codes $u(x) \equiv a$ for a constant vector a, the centered origin is a global minimizer of L_{TS} with value I_T. In the full space of input-dependent code functions, however,*

$$\frac{d}{d\alpha} L_{\mathrm{TS}}(u_\alpha)\Big|_{\alpha=0} = -\mathbb{E}_X J(T_X, \bar{T}) < 0.$$

Therefore, for any Teacher–Student weight $\lambda_{\mathrm{TS}} > 0$ and any finite quadratic mean-prior coefficient, the combined Teacher–Student plus quadratic mean cost still has a strictly negative right derivative at the origin.

Proof. Any input-independent code induces a constant witness distribution q, so $L_{\mathrm{TS}} = I_T + \mathrm{KL}(\bar{T}\|q) \ge I_T$, with equality at the centered witness $S(0) = \bar{T}$. Proposition 2 gives $G'(0) = \mathbb{E}J(T_X, \bar{T})$ and hence the displayed negative derivative of $L_{\mathrm{TS}} = I_T - G$. The quadratic mean-prior cost is $O(\alpha^2)$ along $u_\alpha = \alpha u_T$, so its right derivative at 0 is zero. □

The function-space result is not merely post-hoc; the remaining question is whether its code-space escape direction survives the encoder parameterization. For a parameterized encoder,

$$\nabla_\phi L_{\mathrm{TS}} = \mathbb{E}[J_\phi u_\phi(x)^\top g_u(x)],$$

where $J_\phi u_\phi(x)$ is the encoder Jacobian and $g_u(x) := \nabla_u \mathrm{KL}(T_x\|S(u))|_{u=u_\phi(x)}$ is the code-space alignment gradient. Thus the encoder Jacobian filters the code-space escape field into parameter space: saturation, small witness-visible singular values, opposing ELBO gradients, or poor Jacobian conditioning can attenuate the realized update. Figure 2 shows the analytic path only; without a uniform lower bound on the relevant Jacobian singular values, no optimizer-independent convergence claim is made.

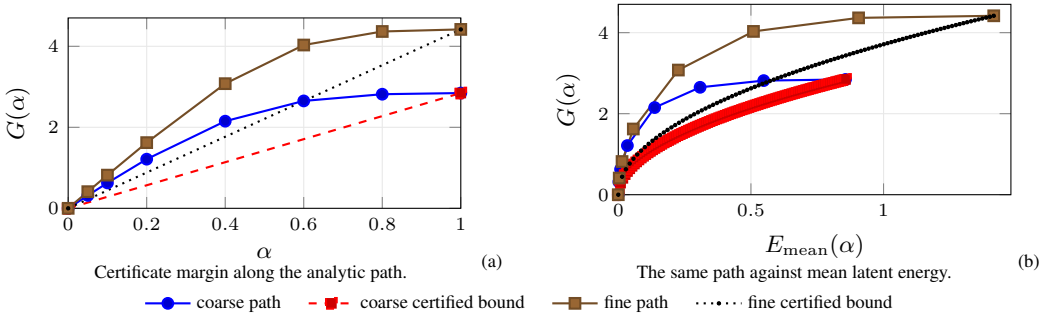

Figure 2: Analytic margin–energy path on CIFAR-100 coarse and fine Teacher views. Left: the exact computed certificate is above the guaranteed linear lower bound αI_T. Right: eliminating α from $E_{\mathrm{mean}}(\alpha) = \alpha^2 E_{\mathrm{mean}}(1)$ yields the certified lower curve $G \ge I_T \sqrt{E_{\mathrm{mean}}/E_{\mathrm{mean}}(1)}$. These are evaluations of the closed-form construction, not learned encoder trajectories or claims of a globally Pareto-optimal frontier.

Preservation and composition. Finally, we ask how much visible drift is safe and how several prescribed views can coexist.

Theorem 2 (Certified visible cylinder)**.** *For the simplex witness, $L_{\mathrm{simp}} := \beta\sqrt{2K/(K-1)}$ satisfies*

$$|\mathrm{KL}(T_x\|S(u')) - \mathrm{KL}(T_x\|S(u))| \le L_{\mathrm{simp}}\|P_{\mathcal{V}}(u'-u)\|.$$

Hence a positive-margin anchor u_{anc} with $G_{\mathrm{anc}} := I_T - L_{\mathrm{TS}}(u_{\mathrm{anc}}) > 0$ remains certified under any refinement whose visible displacement is uniformly below $G_{\mathrm{anc}}/L_{\mathrm{simp}}$; witness-null displacement is unrestricted. At u_α, for any $0 < \gamma < 1$, choosing visible radius $\varepsilon_\alpha = \gamma\alpha I_T/L_{\mathrm{simp}}$ guarantees $G \ge (1-\gamma)\alpha I_T > 0$.

Proposition 3 (Orthogonal multi-view composition)**.** *For view m, write K_m for its number of classes; V_m and $u_{T,m}$ for its witness matrix and analytic Teacher code; and $I_m, \mathcal{E}_m, \beta_m, G_m$ for its Teacher information, log-ratio energy, witness gain, and certificate margin. For M Teachers placed in pairwise orthogonal visible blocks $\mathcal{V}_m$ of dimensions $K_m - 1$, the code*

$$u(x) = \sum_{m=1}^{M} \alpha_m u_{T,m}(x) + h(x), \qquad h(x) \in \bigcap_m \ker V_m, \tag{4}$$

obeys $G_m \ge \alpha_m I_m$ simultaneously, with additive visible energy $\sum_m c_m\alpha_m^2$, $c_m = (K_m - 1)\mathcal{E}_m/(2K_m\beta_m^2)$. Thus latent dimension $d_z \ge \sum_m (K_m - 1)$ is a sufficient capacity rule for orthogonally allocated categorical views, and requested margins τ_m are achieved constructively by $\alpha_m = \tau_m/I_m$. Dependent or hierarchical views may admit shared geometries; orthogonality is the clean sufficient construction studied here.

Corollary 3 (Requested-margin energy budget)**.** *For requested margins $0 \le \tau_m \le I_m$, choosing $\alpha_m = \tau_m/I_m$ guarantees all $G_m \ge \tau_m$ and uses visible mean energy $\sum_m c_m(\tau_m/I_m)^2$.*

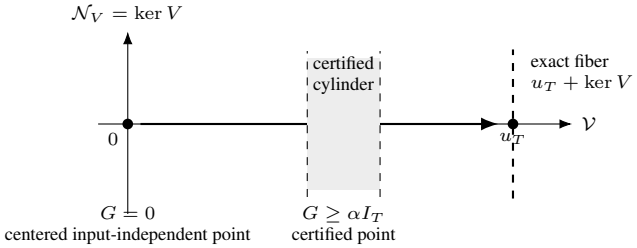

Figure 3: One geometry supports realization, routing, movement, and preservation. The analytic ray lies in the witness-visible space; all exact Teacher realizations share the affine witness-null fiber. Around a positive-margin anchor, only visible displacement is bounded.

Training objective. The soft RST objective is defined in the Introduction. A routed variant may instead use $u_\phi(x) = \alpha_\star u_T(x) + Nh_\phi(x)$ with $VN = 0$, leaving the common witness-null component learnable. This hard parameterization and the analytic code $u_T(x)$ are constructions, not assumptions on ordinary soft-objective runs; Teacher construction and implementation details are in the appendix.

3 Experimental evidence

We organize the evidence by the claim it tests. All reported certificates use frozen Teachers and witnesses restricted to the declared representation. The broad single-view studies use the PCA–GMM Teacher construction described in the appendix; the CIFAR-100 coarse/fine study uses label-defined Teachers specifically to isolate routing and composition rather than Teacher discovery.

Table 2: Five-dataset boundary crossing (mean $\pm$ sample s.d. over five seeds). Two fixed RST variants are shown for every dataset, avoiding per-dataset best-variant selection. $G_T > 0$ is sufficient; $G_T \le 0$ is inconclusive.

Dataset	VAE G_T	RST-α-prefit G_T	RST-α-logit-prefit G_T
Fashion-MNIST	-0.277 ± 0.328	$\mathbf{1.503 \pm 0.060}$	$\mathbf{1.626 \pm 0.054}$
CIFAR-10	-1.987 ± 0.360	$\mathbf{1.841 \pm 0.062}$	$\mathbf{1.836 \pm 0.041}$
CIFAR-100	-2.074 ± 0.242	$\mathbf{1.488 \pm 0.056}$	$\mathbf{1.542 \pm 0.039}$
SVHN	-1.088 ± 0.133	$\mathbf{1.943 \pm 0.026}$	$\mathbf{1.967 \pm 0.029}$
Tiny-ImageNet-200	-2.279 ± 0.360	$\mathbf{1.321 \pm 0.036}$	$\mathbf{1.317 \pm 0.101}$

Boundary crossing, stress tests, escape, and preservation. Across five image datasets, vanilla VAE is certificate-negative while both fixed RST variants in Table 2 are positive in every case; stronger KL pressure preserves the sign separation. Standard warm-up, cyclical annealing, and free-bits controls remain certificate-negative, whereas witness-null routing preserves the predicted logits while leaving reconstruction comparable. Teacher degradation lowers the margin, shuffled Teachers are weaker, and random Dirichlet Teachers are near zero, confirming that the certificate tracks the *specified* Teacher rather than semantic correctness. Starting in the collapse neighborhood, escape runs reach $G_T \approx 2.0$ on CIFAR-10, 1.6–1.7 on CIFAR-100, and 1.31–1.43 on Tiny-ImageNet-200; encoder-gradient diagnostics show the nonzero code-space field survives the parameterization at practical scale. Under strong-KL refinement, an unconstrained route has safety-slack margin $G_\tau := G_T - \tau = -0.671$ and 0 active units (AU), whereas certified refinement maintains a minimum $G_\tau = 2.73490$ and AU$= 64$. The closed-form CIFAR-100 path is numerically exact to floating-point precision. Full breadth statistics are in Appendix D.1.2.

Activity, decodability, and prescribed alignment. Table 3 summarizes matched CIFAR-100 controls with the same coarse/fine Teachers and 55-epoch setup. The fixed-variance δ model is fully active yet fails both declared-view certificates. More sharply, Aux-full and Aux-block learn the Teachers without using the simplex witness during training; after freezing each encoder, a fresh linear probe achieves positive Teacher-relative margins while the predeclared simplex margins remain negative in every seed. RST is positive under both readouts: the information is present in Aux codes but not bound to the declared interface.

Table 3: CIFAR-100 matched controls (five seeds). G_{pre} uses the predeclared simplex; G_{probe} fits a linear probe only after freezing the encoder. A dash means the probe was not part of this diagnostic.

Method	AU	$G_{\mathrm{pre}}^{c/f}$	$G_{\mathrm{probe}}^{c/f}$
VAE	167.8 ± 2.8	-3.344 ± 1.131 / -4.162 ± 0.396	–
InfoVAE	172.2 ± 10.3	-3.435 ± 0.913 / -4.650 ± 0.340	–
Fixed-var. δ	192.0 ± 0.0	-4.718 ± 0.368 / -5.944 ± 0.193	–
Aux-full	192.0 ± 0.0	-3.089 ± 0.779 / -4.892 ± 0.377	$\mathbf{1.078 \pm 0.013}$ / $\mathbf{1.710 \pm 0.028}$
Aux-block	186.8 ± 2.2	-2.697 ± 1.059 / -5.136 ± 0.758	$\mathbf{0.986 \pm 0.013}$ / $\mathbf{1.619 \pm 0.037}$
RST-joint	191.8 ± 0.4	$\mathbf{1.057 \pm 0.027}$ / $\mathbf{1.477 \pm 0.026}$	$\mathbf{1.162 \pm 0.016}$ / $\mathbf{1.957 \pm 0.026}$

No-refit reuse across seeds and architectures. Within the ConvVAE Aux-block control, native learned-head accuracy is $36.57/21.88\%$ (coarse/fine), but cross-seed no-adapter reuse falls to $5.20/0.94\%$; Procrustes makes $39/40$ off-diagonal margins positive and ridge-affine makes $40/40$. The fixed RST witness directly reads every Conv-RST seed. Table 4 then changes the producer: five ResVAE RST models use the same frozen Teachers, fixed witness, and $d_z = 192$ block allocation, without regression to $u_T(x)$ or hard routing. A predeclared 45k/5k train/validation split selects checkpoints by the fixed validation objective; the official test is accessed only after selection.

Table 4: Cross-architecture fixed-interface audit on CIFAR-100 (five seeds). "Fixed acc." uses the same predeclared witness; "native Aux" uses the jointly trained control head. Conv checkpoints predate the 45k/5k Res split, so absolute backbone performance is not compared.

Producer	Method	$G_{\mathrm{pre}}^{c/f}$	fixed acc. c/f (%)	native Aux acc. c/f (%)
ConvVAE	RST	$\mathbf{1.057 \pm 0.027}$ / $\mathbf{1.477 \pm 0.026}$	$\mathbf{43.84 \pm 1.30}$ / $\mathbf{33.58 \pm 0.56}$	–
ResVAE	RST	$\mathbf{0.869 \pm 0.057}$ / $\mathbf{1.083 \pm 0.105}$	$\mathbf{37.28 \pm 1.94}$ / $\mathbf{23.87 \pm 2.20}$	–
ConvVAE	Aux-block	-2.697 ± 1.059 / -5.136 ± 0.758	4.77 ± 0.71 / 1.01 ± 0.17	36.57 ± 0.85 / 21.88 ± 2.24
ResVAE	Aux-block	-2.543 ± 1.004 / -5.314 ± 0.883	4.90 ± 0.75 / 0.85 ± 0.18	20.03 ± 1.11 / 10.38 ± 1.09

Cross-architecture learned-reader transfer. Swapping learned Aux heads across backbones is near chance without refitting: Conv→Res gives $5.11/0.91\%$ and Res→Conv $4.98/0.96\%$ coarse/fine accuracy, with $98/100$ pairwise margins negative. Adapters are fitted on 20,000 training examples only. Procrustes makes $89/100$ margins positive, whereas ridge-affine makes all $100/100$ positive and reaches $21.81/14.74\%$ (Conv→Res) and $24.43/10.47\%$ (Res→Conv). Because Procrustes is incomplete, the mismatch is not explained by a pure rotation. Full per-direction margins, aggregation, and protocol are in Appendix C.2.2.

Controlled interface audit. Two interventions connect the functional result back to the exact geometry. Rotating an RST visible block while keeping the original witness fixed drops coarse/fine accuracy to $4.81/1.00\%$; co-rotating the readout restores $43.84/33.58\%$ with maximum logit error below 5×10^{-6}. Zeroing either visible block returns only that route to the input-independent boundary ($|G_T| < 5 \times 10^{-7}$) while the orthogonal route is numerically unchanged; swapping a block makes the target interface follow the donor while the non-target interface follows the recipient. These are latent-interface audits, not decoder-causality claims; full per-seed results are in Appendix C.2.3.

Selective multi-view routing and composition. CIFAR-100 provides a 20-way coarse and 100-way fine label view. We smooth each label Teacher with $\epsilon = 0.02$ and allocate simplex-visible blocks of dimensions 19 and 99 inside $d_z = 192$, leaving a 74-dimensional common witness-null complement. Five seeds give the selective pattern

$$\boxed{\text{baseline } (-,-), \quad \text{coarse } (+,-), \quad \text{fine } (-,+), \quad \text{joint } (+,+),}$$

with joint-null also $(+,+)$ in every seed. Table 5 gives the mean margins and route accuracies. The experiment intervenes on preassigned routes; it does not test unsupervised semantic discovery.

Table 5: Five-seed selective routing/composition on CIFAR-100. Margins are mean $\pm$ sample s.d.; accuracies are five-seed means.

Method	G_c	G_f	coarse acc.	fine acc.
VAE	-1.660 ± 0.964	-1.728 ± 0.324	5.45%	0.85%
Coarse only	$\mathbf{0.619 \pm 0.017}$	-2.519 ± 0.416	34.54%	0.88%
Fine only	-2.302 ± 1.167	$\mathbf{0.579 \pm 0.013}$	5.36%	20.99%
Joint	$\mathbf{0.686 \pm 0.019}$	$\mathbf{0.655 \pm 0.019}$	36.94%	22.54%
Joint + witness-null	$\mathbf{0.693 \pm 0.017}$	$\mathbf{0.632 \pm 0.014}$	37.70%	23.59%

4 Related work

Posterior collapse and information objectives. Collapse has been linked to decoder dominance or latent underuse (Bowman et al., 2016; Chen et al., 2017), inference lag and optimization (He et al., 2019; Dai et al., 2020), linear-Gaussian structure (Lucas et al., 2019), oversmoothing (Takida et al., 2021), non-identifiability (Wang et al., 2021), and explicit structural or threshold analyses (Dang et al., 2024; Ichikawa and Hukushima, 2024; Hirn, 2026; Li et al., 2026). Remedies alter KL schedules or free bits (Fu et al., 2019; Kingma et al., 2016), inference, posterior constraints, skip paths, decoder regularity, local reconstruction, or reparameterization noise (He et al., 2019; Razavi et al., 2019; Zhao et al., 2019; Tolstikhin et al., 2018; Dieng et al., 2019; Kinoshita et al., 2023; Song et al., 2026; Demisse, 2026). These methods target useful latent information in general. RST instead fixes a particular Teacher-relative information route. Information-theoretic objectives and variational information bounds (Tishby et al., 1999; van den Oord et al., 2018; Poole et al., 2019; Barber and Agakov, 2003) supply related tools; the earlier Teacher-guided certificate (Zhang et al., 2026) established the best-constant/input-independent boundary used here.

Fixed readouts, probing, and compatible representations. Linear probes ask what a separately fitted readout can recover from frozen features (Alain and Bengio, 2016; Hewitt and Liang, 2019). Neural collapse identifies simplex-ETF geometry in terminal supervised representations (Papyan et al., 2020). Separately, fixed-classifier work freezes the terminal readout and lets the feature extractor adapt, including generic fixed classifiers, regular-polytopal/simplex or ETF frames, and hierarchy-aware frames (Hoffer et al., 2018; Pernici et al., 2022; Yang et al., 2022; Liang and Davis, 2023). Compatible-representation work studies direct or transformed reuse across independently trained or updated models (Shen et al., 2020; Gygli et al., 2021; Biondi et al., 2024; Ricci et al., 2025; Biondi et al., 2026), whereas model stitching explicitly learns a connector between frozen networks (Bansal et al., 2021). Concept-bottleneck models likewise make an intermediate concept interface explicit and intervenable (Koh et al., 2020). RST does *not* claim fixed simplex classifiers, stationary/compatible representations, or adapter-free reuse as new. Its distinct object is a Teacher-relative certificate inside a VAE: a frozen possibly soft Teacher specifies the view, a predeclared witness specifies the audited interface, and the same geometry yields an attainable boundary, exact realization/fiber, margin–energy movement, preservation, and multi-view composition.

Simplex geometry. Regular simplices and tight frames are classical (Benedetto and Fickus, 2003). The contribution is not the simplex itself, but the RST-specific chain obtained once that classical geometry is coupled to the Teacher-relative certificate: minimum visible dimension, closed-form soft-Teacher inverse, complete witness-null fiber, certified movement/energy, escape, preservation, and composition. Under the normalized minimal-interface criterion stated in Appendix B.3, the classical regular simplex is the canonical realization up to rotation.

5 Discussion and scope

What prescribed alignment adds. If a downstream task may refit an arbitrary probe, ordinary decodability can suffice; the new readout can absorb coordinate changes. RST addresses the narrower case in which route identity and readout are part of the specification. The formal rotation counterexample (Appendix C.1) and the cross-seed/cross-architecture transfer studies show why these notions differ: Aux information is present and recoverable, yet direct producer–consumer reuse fails without coordinate adaptation, while one predeclared RST witness remains readable across two producer backbones. This is a certificate-level interface claim, not a claim of intrinsic semantic coordinates, universal architecture invariance, or a first compatible-representation method. The functional contrast is external specification versus post-hoc negotiation: RST fixes the consumer geometry before producer training, whereas Aux lets the producer and learned head co-adapt. Thus the cross-architecture result tests portability of a fixed contract for the tested ConvVAE/ResVAE pair, not representation identity. The preservation theorem sharpens this observation: once the realized code lies inside the same certified visible cylinder, the guarantee no longer depends on the producer parameterization.

Corollary 4 (Cross-architecture prescribed-interface realizability)**.** *Fix the Teacher, simplex witness, evaluation distribution, and* $0 < \alpha \leq 1$. *For any encoder architecture* a, *let* $u^{(a)}(x)$ *be its realized audited code and* $G_T^{(a)}$ *its corresponding certificate margin, and define* $\varepsilon_a := \sup_x \|P_{\mathcal{V}}(u^{(a)}(x) - u_\alpha(x))\|$ *on the audited support. Then* $G_T^{(a)} \geq \alpha I_T - L_{\mathrm{simp}}\varepsilon_a$; *hence* $\varepsilon_a < \alpha I_T / L_{\mathrm{simp}}$ *certifies the same predeclared interface for every such producer. If* $u^{(a)} = u_T + h_a$ *with* $h_a \in \ker V$, *then* $G_T^{(a)} = I_T$. *This is a statement about realized code functions, not optimizer-independent convergence.*

Why a certificate rather than another compatibility objective? Compatibility objectives can encourage direct reuse, but reuse alone does not quantify which frozen Teacher information crosses which declared route. RST separates specification from training: the Teacher, witness, and reference information I_T are fixed before the audited optimization, so $G_T > 0$ has an information-theoretic meaning that does not depend on fitting a downstream adapter. The construction then turns the same certificate into an explicit interface contract—what must be visible, what may vary in the witness-null complement, how a positive state may move while remaining certified, and how several declared contracts compose. The compatibility-style experiment above is evidence for this contract, not the definition of the contribution.

Scope. The guarantee is relative to the frozen Teacher; unsupervised Teacher discovery is separate. $G_T \leq 0$ is inconclusive, re-instantiation requires the stated Markov condition, and the analytic ray/escape field are function-space statements rather than optimizer-independent convergence guarantees. Preservation begins from a positive-margin anchor, and neither witness-null invariance nor certification proves decoder use. RST therefore claims a sufficient, Teacher-relative, route-specific information guarantee—not semantic completeness or universal collapse resolution.

6 Conclusion

RST turns an established Teacher-relative input-independent boundary into a constructive prescribed-information interface. One certificate/geometry supports exact realization, residual freedom, movement and cost, escape, preservation, composition, architecture-independent certification of realized interfaces, and—empirically—no-refit reuse across two producer backbones under one predeclared witness. The novelty is therefore neither the classical simplex nor compatibility as a problem, but a certifiable bridge from a declared (possibly soft) information view to a constructive, reusable VAE interface.

Reproducibility Statement

The theorem statements specify the assumptions needed for the certificate, simplex realization, path, escape, preservation, and composition results. The five-seed CIFAR-100 studies, including auxiliary-supervision and cross-architecture controls, archive run arguments, histories, result summaries, and system metadata. The appendix records Teacher construction, architecture and optimization settings, expanded diagnostics, numerical audits, and additional derivations needed to reproduce the reported procedures.

AI Use Statement

Generative AI tools were used to assist with conceptual discussion, mathematical formulation and proof checking, experimental design, code implementation and debugging, interpretation of experimental outputs, Chinese–English translation, literature search and organization, and drafting and editing portions of the manuscript. AI-assisted material was reviewed by the authors, who made the final scientific decisions and take responsibility for the mathematical claims, experiments, results, citations, code, and text in this work.

REFERENCES

David Barber and Felix Agakov. The IM Algorithm: A Variational Approach to Information Maximization. NeurIPS, 2003.

John J. Benedetto and Matthew Fickus. Finite Normalized Tight Frames. Advances in Computational Mathematics, 18:357–385, 2003.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. CoNLL, 2016.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational Lossy Autoencoder. ICLR, 2017.

Bin Dai, Ziyu Wang, and David Wipf. The Usual Suspects? Reassessing Blame for VAE Posterior Collapse. ICML, 2020.

Hien Dang, Tho-Huu Tran, Tan Nguyen, and Nhat Ho. Beyond Vanilla Variational Autoencoders: Detecting Posterior Collapse in Conditional and Hierarchical Variational Autoencoders. ICLR, 2024.

Girum Demisse. λ-VAE: Variance Equalization for Posterior Collapse. arXiv:2607.05531, 2026.

Adji B. Dieng, Yoon Kim, Alexander M. Rush, and David M. Blei. Avoiding Latent Variable Collapse with Generative Skip Models. AISTATS, 2019.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing. NAACL, 2019.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging Inference Networks and Posterior Collapse in Variational Autoencoders. ICLR, 2019.

Johannes Hirn. Posterior Collapse as Automatic Spectral Pruning. arXiv:2605.22691, 2026.

Yuma Ichikawa and Koji Hukushima. Learning Dynamics in Linear VAE: Posterior Collapse Threshold, Superfluous Latent Space Pitfalls, and Speedup with KL Annealing. AISTATS, 2024.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. ICLR, 2014.

Yuri Kinoshita, Kenta Oono, Kenji Fukumizu, Yuichi Yoshida, and Shin-Ichi Maeda. Controlling Posterior Collapse by an Inverse Lipschitz Constraint on the Decoder Network. ICML, 2023.

Zhen Li, Fan Zhang, Zheng Zhang, and Yu Chen. Posterior Collapse as a Phase Transition in Variational Autoencoders. Physica A: Statistical Mechanics and its Applications, 683:131228, 2026.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Understanding Posterior Collapse in Generative Latent Variable Models. Deep Generative Models for Highly Structured Data Workshop at ICLR, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748, 2018.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On Variational Bounds of Mutual Information. ICML, 2019.

Ali Razavi, Aaron van den Oord, Ben Poole, and Oriol Vinyals. Preventing Posterior Collapse with delta-VAEs. ICLR, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. ICML, 2014.

Hyunsoo Song, Seungwhan Kim, and Seungkyu Lee. Toward Architecture-Agnostic Local Control of Posterior Collapse in Variational Auto-encoders. SSRN preprint 7190902, 2026.

Yuhta Takida, Wei-Hsiang Liao, Toshimitsu Uesaka, Shusuke Takahashi, and Yuki Mitsufuji. Preventing Posterior Collapse Induced by Oversmoothing in Gaussian VAE. arXiv:2102.08663, 2021.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein Auto-Encoders. ICLR, 2018.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The Information Bottleneck Method. Allerton Conference on Communication, Control, and Computing, 1999.

Yixin Wang, David M. Blei, and John P. Cunningham. Posterior Collapse and Latent Variable Non-identifiability. NeurIPS, 2021.

Zegu Zhang, Jianhua Peng, and Jian Zhang. A Testable Certificate for Constant Collapse in Teacher-Guided VAEs. arXiv:2605.05813, 2026.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Balancing Learning and Inference in Variational Autoencoders. AAAI, 33(01):5885–5892, 2019. doi:10.1609/aaai.v33i01.33015885.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. arXiv:1610.01644, 2016.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *NeurIPS*, 34:225–236, 2021.

Niccolò Biondi, Federico Pernici, Simone Ricci, and Alberto Del Bimbo. Stationary representations: Optimally approximating compatibility and implications for improved model replacements. *CVPR*, pp. 28793–28804, 2024.

Niccolò Biondi, Federico Pernici, Simone Ricci, and Alberto Del Bimbo. A stationary (and therefore compatible) representation is all you need. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. Online ahead of print; doi:10.1109/TPAMI.2026.3692279.

Michael Gygli, Jasper Uijlings, and Vittorio Ferrari. Towards reusable network components by learning compatible representations. *AAAI*, 35(9):7620–7629, 2021. doi:10.1609/aaai.v35i9.16932.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *EMNLP-IJCNLP*, pp. 2733–2743, 2019. doi:10.18653/v1/D19-1275.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. NIPS Deep Learning and Representation Learning Workshop, 2015. arXiv:1503.02531.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Fix your classifier: the marginal value of training the last weight layer. *ICLR*, 2018.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *NeurIPS*, pp. 4743–4751, 2016.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. *ICML*, pp. 5338–5348, 2020.

Tong Liang and Jim Davis. Inducing neural collapse to a fixed hierarchy-aware frame for reducing mistake severity. *ICCV*, pp. 1443–1452, 2023. doi:10.1109/ICCV51070.2023.00139.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. doi:10.1073/pnas.2015509117.

Federico Pernici, Matteo Bruni, Claudio Baecchi, and Alberto Del Bimbo. Regular polytope networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4373–4387, 2022. doi:10.1109/TNNLS.2021.3056762.

Simone Ricci, Niccolò Biondi, Federico Pernici, Ioannis Patras, and Alberto Del Bimbo. λ-orthogonality regularization for compatible representation learning. *NeurIPS*, 38:32841–32868, 2025. doi:10.52202/085713-0974.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi:10.1007/BF02289451.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. *CVPR*, pp. 6368–6377, 2020.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *NeurIPS*, 35:37991–38002, 2022.

APPENDIX

A READER MAP, NOTATION, AND CLAIM–EVIDENCE MAP

This opening block is a lookup and navigation layer. It defines notation once, maps each main claim to its proof/evidence, and points to the five evidence blocks below.

NOTATION AND TERMINOLOGY

The main text follows a definition-before-use convention for nonstandard symbols. The tables below are a lookup aid; symbols used only inside a specialized appendix derivation are defined locally there.

RST name and core probabilistic notation. RST always expands to **Reconstruction–Student–Teacher**: R is the ordinary VAE reconstruction branch, S is the fixed Student witness/readout, and T is the frozen Teacher specification. R is never used as the mathematical representation variable. This Teacher/Student terminology differs from knowledge distillation (Hinton et al., 2015): here the Teacher is a frozen information specification and the Student is a fixed code-only witness, not a learned predictor compressed from another model.

Symbol / term	Name	Meaning in this paper
X, x	input random variable / realization	Data sample and a realized input.
T_X, T_x	frozen Teacher view	$T_X := T(\cdot \mid X)$ and $T_x := T(\cdot \mid x)$; a full-support categorical distribution over the declared view.
C	Teacher variable	Auxiliary categorical variable sampled as $C \mid X \sim T_X$.
$\bar{T}$	Teacher marginal	$\mathbb{E}_X[T_X]$; the optimal constant prediction for the declared Teacher view.
Z, Z_m	audited representation	Deterministic or stochastic representation/statistic being audited; distinct from R in RST.
Z_μ	mean representation	$Z_\mu := u_\phi(X)$; the deterministic encoder mean audited in the main experiments.
z	sampled VAE latent	$z \sim q_\phi(z \mid x)$; the sample passed to the decoder $p_\theta(x \mid z)$. Lowercase z is not used for the deterministic witness code.
$S(\cdot \mid Z)$	Student / fixed witness	Predeclared code-only readout used to audit the chosen representation interface.
I_T	Teacher information / boundary	$I_T := I(C;X) = \mathbb{E}_X \mathrm{KL}(T_X \Vert \bar{T})$; the attainable input-independent loss boundary for the centered witness.
L_{TS}	Teacher–Student loss	$\mathbb{E}_{X,Z} \mathrm{KL}(T_X \Vert S(\cdot \mid Z))$.
G_T	certificate margin	$G_T := I_T - L_{\mathrm{TS}}$; $G_T > 0$ implies $I(C;Z) \geq G_T > 0$.
input-independent reference	audited reference class	$Z \perp X$; every such representation obeys $L_{\mathrm{TS}} \geq I_T$.
constant collapse	deterministic special case	$Z \equiv u_0$, hence input-independent.
prescribed alignment	fixed-interface requirement	The declared Teacher view must be readable at the predeclared representation/readout interface; $G_T > 0$ is a sufficient certificate, not the definition itself.
learned decodability	post-hoc readability	The view can be recovered after fitting a new head; the readout coordinates may change.
adapter-free interface	functional interpretation	Evaluation or reuse with the predeclared witness, without refitting an adapter/head.

Simplex, optimization, and diagnostic notation.

Table 6: Simplex, optimization, and diagnostic notation.

Symbol	Name	Meaning in this paper
K	number of Teacher classes	A K-class full-support categorical Teacher needs $K-1$ simplex-visible dimensions.
v_k, V, β	simplex witness geometry	v_k are unit regular-simplex vertices, V stacks them as rows, and $\beta > 0$ is the fixed witness gain.
u	deterministic witness code	Generic deterministic input to the simplex witness $S(u) = \mathrm{softmax}(\beta V u + \log \bar{T})$.
$\mathcal{V}, P_{\mathcal{V}}$	witness-visible space / projector	$\mathcal{V} = \mathrm{row}(V)$; $P_{\mathcal{V}}$ is its orthogonal projector.
$\mathcal{N}_V$	witness-null space	$\mathcal{N}_V = \ker V = \mathcal{V}^\perp$; geometric directions invisible to the witness, not the same as $Z \perp X$.
$r(x)$	centered log-ratio vector	Centered coordinates of $\log(T_{x,k}/\bar{T}_k)$ used by the analytic inverse.
$u_T(x)$	analytic Teacher code	Unique minimum-norm visible code satisfying $S(u_T(x)) = T_x$; all exact solutions are $u_T(x) + \ker V$.
$\mathcal{E}_T$	Teacher log-ratio energy	$\mathbb{E}\Vert r(X)\Vert^2$; determines the closed-form mean-prior cost.
u_α, α	analytic Teacher path	$u_\alpha(x) = \alpha u_T(x)$, $0 \leq \alpha \leq 1$.
$G(\alpha)$	path margin	$I_T - L_{\mathrm{TS}}(u_\alpha)$.
$E_{\mathrm{mean}}(\alpha)$	mean-prior energy	$\frac{1}{2}\mathbb{E}\Vert u_\alpha(X)\Vert^2$.
N, h_ϕ	witness-null residual parameterization	Columns of N span $\ker V$; $N h_\phi(x)$ can vary without changing the fixed witness logits.

$\mathcal{L}_{\mathrm{rec}}, \mathcal{L}_{\mathrm{KL}}$	standard VAE terms	Reconstruction loss and posterior-to-prior KL; weights are β_{KL} and λ_{TS} for KL and Teacher–Student alignment.
d_z	latent dimension	Dimension of the VAE latent/mean code used in an experiment.
G_τ	safety-slack margin	$G_\tau := G_T - \tau$ when a reserved positive slack τ is imposed in preservation tests.
G_c, G_f	coarse / fine margins	CIFAR-100 certificate margins for the 20-way coarse and 100-way fine Teacher views.
AU	active units	Generic latent-activity diagnostic; not equivalent to prescribed alignment.
$R_{\hat{x}}$	reconstruction-side Teacher posterior	Frozen Teacher evaluated on the reconstruction $\hat{x}$; an optional decoder-side diagnostic object.
L_{TR}, L_{SR}	reconstruction-side diagnostics	Teacher–Reconstruction and Student–Reconstruction KL diagnostics; neither enters G_T.

APPENDIX ROADMAP

The appendix is evidence-first and split into six blocks. Appendix A provides notation and a claim–evidence map. Appendix B places the core proofs and constructive geometry in dependency order. Appendix C joins the exact decodability–alignment separation with the matched interface experiments. Appendix D contains full empirical tables and reproducibility details. Appendix E collects mechanistic and numerical audits. Appendix F states route re-instantiations and limitations, with prospective extensions explicitly separated from validated claims.

CLAIM–EVIDENCE MAP

The table separates exact mathematical claims from the experiment that probes each consequence. The analytic path audit checks implementation of closed-form identities; it is not treated as evidence that SGD follows the analytic ray.

Main claim	Formal support	Empirical check / role
G_T lower-bounds prescribed information	Thm. 1	Positive margins on five datasets
I_T is the exact attainable input-independent boundary for the centered RST witness	Thm. 1	Constant/input-independent controls attain the boundary; RST crosses it
The certificate re-instantiates for a chosen route/statistic	Cor. 1	Block-wise studies; log-variance cross-statistic sanity check
$K-1$ dimensions are minimal; simplex is normalized minimax-stable	Props. 1, 1	Geometric statement; no learned-witness premise
Exact Teacher solutions form $u_T + \ker(V)$ and u_T is minimum norm	Prop. 1	Closed-form inverse numerical audit; routed runs
Analytic path has concave margin and quadratic energy	Prop. 2	Analytic–computation audit only (Fig. 2)
Witness-null motion is invariant; positive anchors have certified visible cylinders	Prop. 1, Thm. 2	Witness-null-route runs; trajectory-mined certified refinement
Multiple orthogonal views compose with additive energy budgets	Prop. 3, Cor. 3	CIFAR-100 coarse/fine selective sign pattern; moving-shape block test
Generic activity does not imply the prescribed certificates	Certificate definition plus matched controls	δ control: $192/192$ AU but both margins negative
Learned decodability does not imply prescribed alignment	Prop. 5 plus fixed-witness definition	Exact rotation counterexample; frozen probes are positive while fixed margins are negative; cross-seed/cross-architecture reuse fails and adapters recover

Fast verification route. Core mathematical support is in Appendix B; interface separation in Appendix C; full empirical evidence in Appendix D; implementation-facing audits in Appendix E; and extensions/scope in Appendix F.

B CORE PROOFS AND CONSTRUCTIVE GEOMETRY

Back to main text: Section 2 (constructive theory)

This block follows the main theoretical dependency: certificate and boundary → canonical simplex realization → exact fiber → path/energy → preservation → multi-view composition. Classical geometric facts are used as ingredients; the claims here are the RST-specific consequences and their proofs.

B.1 CERTIFICATE, BOUNDARY, AND CORE CONSTRUCTION PROOFS

This section records the proofs that are only stated compactly in the main text. It is deliberately separate from the later auxiliary identities and numerical audits.

Theorem 1 (information decomposition and input-independent boundary). Because $C - X - Z$, $P(C \mid X, Z) = P(C \mid X) = T(C \mid X)$. Insert $P(C \mid Z)$ into the alignment loss:

$$\begin{aligned} L_{\mathrm{TS}} &= \mathbb{E} \log \frac{P(C \mid X, Z)}{S(C \mid Z)} \\ &= \mathbb{E} \log \frac{P(C \mid X, Z)}{P(C \mid Z)} + \mathbb{E} \log \frac{P(C \mid Z)}{S(C \mid Z)} \\ &= I(C; X \mid Z) + \mathbb{E}_Z \mathrm{KL}(P(C \mid Z) \| S(C \mid Z)). \end{aligned}$$

Also $I_T = I(C; X)$. Since $I(C; Z \mid X) = 0$, the chain rule gives $I(C; X) = I(C; Z) + I(C; X \mid Z)$. Subtracting the displayed identity from I_T yields

$$G_T = I(C; Z) - \mathbb{E}_Z \mathrm{KL}(P(C \mid Z) \| S(C \mid Z)) \le I(C; Z).$$

If $Z \perp X$, then $Z \perp C$ and $P(C \mid Z) = \bar{T}$, so

$$L_{\mathrm{TS}} = I_T + \mathbb{E}_Z \mathrm{KL}(\bar{T} \| S(C \mid Z)) \ge I_T.$$

For the centered RST witness, the constant code $u = 0$ satisfies $S(\cdot \mid 0) = \bar{T}$, hence the boundary is attained.

Corollary 1 (route-local re-instantiation). Replace (C, Z, T, S) by $(C_m, Z_m, T^{(m)}, S_m)$ in the preceding argument. No step uses a special property of the posterior mean; only the stated Markov relation and witness support condition are required.

Proposition 1 (minimal realization, exact fiber, and witness-null invariance). Softmax is invariant to a common logit shift, so identifiable categorical logits lie in $\mathbf{1}^\perp$, a $(K-1)$-dimensional space. A linear-softmax witness that represents an open set of full-support K-class distributions therefore needs visible rank at least $K - 1$. The regular simplex has exactly this rank and satisfies

$$V V^\top = \frac{K}{K-1} \left(I - \frac{1}{K} \mathbf{1}\mathbf{1}^\top \right), \qquad V^\top V = \frac{K}{K-1} P_{\mathcal{V}}.$$

For the centered log-ratio vector $r(x) \in \mathbf{1}^\perp$, the definition $u_T(x) = \frac{K-1}{K\beta} V^\top r(x)$ gives $\beta V u_T(x) = r(x)$ and hence $S(u_T(x)) = T_x$. Conversely, if $S(u) = T_x$, centering the equal logits removes the common shift and gives $V(u - u_T) = 0$, so the full solution set is $u_T + \ker V$. Because $u_T \in \mathcal{V}$ and $\ker V = \mathcal{V}^\perp$, u_T is the unique minimum-norm realization. Finally, $Vh = 0$ for every $h \in \ker V$, so $S(u + h) = S(u)$ and the certificate is exactly invariant to witness-null motion.

Proposition 2 (analytic margin–energy path). For one sample define $D_x = \mathrm{KL}(T_x \| \bar{T})$, $a_k(x) = \log(T_{x,k} / \bar{T}_k)$, and $A_x(\alpha) = \log \sum_k \bar{T}_k e^{\alpha a_k(x)}$. Along $u_\alpha = \alpha u_T$,

$$\ell_x(\alpha) := \mathrm{KL}(T_x \| S(u_\alpha)) = (1 - \alpha) D_x + A_x(\alpha), \qquad \ell_x''(\alpha) = \mathrm{Var}_{S(u_\alpha)}\, a(x) \ge 0.$$

Thus $L_{\mathrm{TS}}(\alpha)$ is convex and nonincreasing, so $G(\alpha) = I_T - L_{\mathrm{TS}}(\alpha)$ is concave and nondecreasing. Since $G(0) = 0$ and $G(1) = I_T$, concavity gives $G(\alpha) \ge \alpha I_T$. Differentiating at the origin yields

$$G'(0) = \mathbb{E}_X \left[\mathrm{KL}(T_X \| \bar{T}) + \mathrm{KL}(\bar{T} \| T_X) \right].$$

The energy identity follows immediately from $u_\alpha = \alpha u_T$: $E_{\mathrm{mean}}(\alpha) = \alpha^2 E_{\mathrm{mean}}(1)$.

Theorem 2 (certified visible cylinder). For $\ell_x(u) = \mathrm{KL}(T_x \| S(u))$,

$$\nabla_u \ell_x(u) = \beta V^\top (S(u) - T_x) \in \mathcal{V}.$$

Since $\|S(u) - T_x\|_2 \le \sqrt{2}$ and $\|V\|_2 = \sqrt{K/(K-1)}$, $\|\nabla_u \ell_x(u)\|_2 \le L_{\mathrm{simp}} := \beta \sqrt{2K/(K-1)}$. The mean-value inequality therefore gives

$$|\ell_x(u') - \ell_x(u)| \le L_{\mathrm{simp}} \|P_{\mathcal{V}}(u' - u)\|_2.$$

Averaging over x yields the stated margin bound. If the displacement lies in $\mathcal{N}_V = \ker V$, the right-hand side is zero exactly.

Corollary 4 (cross-architecture prescribed-interface realizability). Apply Theorem 2 pointwise with anchor $u_\alpha(x)$ and representation $u^{(a)}(x)$, then average over the common evaluation distribution. Proposition 2 gives $G(u_\alpha) \ge \alpha I_T$, hence $G_T^{(a)} \ge \alpha I_T - L_{\mathrm{simp}} \varepsilon_a$. If $u^{(a)} = u_T + h_a$ with $h_a \in \ker V$, Proposition 1 gives exact witness invariance and therefore $L_{\mathrm{TS}} = 0$, so $G_T^{(a)} = I_T$. No property of the internal parameterization is used beyond realization of the stated code function.

Proposition 3 and Corollary 3 (orthogonal composition). Pairwise orthogonality makes witness m insensitive to every other visible block and to the common witness-null complement. Applying the single-view path bound blockwise gives $G_m \geq \alpha_m I_m$ simultaneously. Pythagoras gives additive visible mean energy $\sum_m c_m \alpha_m^2$. Setting $\alpha_m = \tau_m / I_m$ therefore achieves every requested margin $G_m \geq \tau_m$ with the stated additive budget.

B.2 Constructing the simplex vertices

The simplex geometry is fixed once K and d_z are chosen. It is not learned. In $\mathbb{R}^K$, let $e_1, \ldots, e_K$ be the standard basis and define

$$u_k = e_k - \frac{1}{K}\mathbf{1}.$$

Then $\sum_k u_k = 0$ and the points lie in the $(K-1)$-dimensional hyperplane orthogonal to $\mathbf{1}$. After normalization,

$$\tilde{v}_k = \frac{u_k}{\|u_k\|},$$

one obtains

$$\tilde{v}_i^\top \tilde{v}_j = \begin{cases} 1, & i = j, \\ -\frac{1}{K-1}, & i \neq j. \end{cases}$$

If $d_z > K-1$, the simplex is embedded into any fixed $(K-1)$-dimensional subspace of $\mathbb{R}^{d_z}$, with the remaining coordinates set to zero. Thus the difficult design choice is the teacher posterior T_x, not the simplex vertices.

B.3 Path, inverse, and certified-cylinder proofs

B.3.1 Tight-frame identities and the minimum-norm inverse

For completeness, the regular-simplex Gram matrix is

$$VV^\top = \frac{K}{K-1}\left(I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right).$$

Its nonzero eigenvalues are all $K/(K-1)$, so the rows form a tight frame for $\mathcal{V}$ and $V^\top V = \frac{K}{K-1} P_{\mathcal{V}}$. These identities simultaneously give the inverse in Proposition 1, its energy, and the Lipschitz constant in Theorem 2. They also show why an arbitrary learned linear witness does not generally retain the same closed-form geometry: unequal singular values destroy the scaled-isometry property.

The normalized optimality can be stated as an equivalence that makes the design choice non-arbitrary.

Proposition 4 (Canonicality under the normalized minimal-interface criterion)**.** *Let $W \in \mathbb{R}^{K \times (K-1)}$ satisfy $W^\top \mathbf{1} = 0$, $\operatorname{rank}(W) = K-1$, and unit row norms. The following are equivalent: (i) $\|W\|_2^2 = K/(K-1)$, the minimum possible spectral norm under these constraints; (ii) all nonzero singular values equal $\sqrt{K/(K-1)}$; (iii) $WW^\top = \frac{K}{K-1}(I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top)$; and (iv) the rows are regular-simplex vertices up to a right orthogonal transformation. Hence the regular simplex is the unique normalized minimal-dimensional witness geometry, up to rotation, that minimizes the worst-case linear sensitivity entering the certified-cylinder radius.*

Proof. The unit-row constraint gives $\|W\|_F^2 = K$, while rank $K-1$ gives $K-1$ nonzero singular values. Therefore their largest square is at least their average $K/(K-1)$, with equality iff all are equal. Centering and full rank make $\mathbf{1}$ the unique left-null direction, so equality forces $WW^\top = \frac{K}{K-1} P_{\mathbf{1}^\perp}$, whose diagonal entries are one and off-diagonal entries are $-1/(K-1)$. This is the regular-simplex Gram matrix. Conversely that Gram matrix has the stated singular values and spectral norm. Right multiplication by an orthogonal matrix preserves the Gram matrix and accounts for the remaining coordinate freedom. □

This classical tight-frame characterization is not claimed as a new frame-theoretic result. Its role here is to show that the same normalized geometry selected by the RST interface criteria is also the geometry that supports the closed-form inverse/fiber and the explicit preservation constant.

B.3.2 Jeffreys slope and concavity of the certificate path

For one sample let $D_x = \mathrm{KL}(T_x \| \bar{T})$ and $a_k = \log(T_{x,k}/\bar{T}_k)$. With

$$A_x(\alpha) = \log \sum_k \bar{T}_k e^{\alpha a_k},$$

we have

$$\ell_x(\alpha) = (1-\alpha)D_x + A_x(\alpha), \qquad \ell_x''(\alpha) = \mathrm{Var}_{q_{\alpha,x}}\, a \geq 0.$$

At the origin,

$$A_x'(0) = \sum_k \bar{T}_k \log \frac{T_{x,k}}{\bar{T}_k} = -\mathrm{KL}(\bar{T}\|T_x),$$

so

$$-\ell_x'(0) = D_x + \mathrm{KL}(\bar{T}\|T_x) = J(T_x, \bar{T}).$$

At $\alpha = 1$, $q_{1,x} = T_x$ and $\ell_x'(1) = 0$. Hence the improvement is steepest at the collapsed endpoint and gradually saturates toward exact Teacher matching.

B.3.3 A CLOSED-FORM CERTIFICATE–ENERGY ACHIEVABILITY BOUND

For a requested margin $0 < \tau < I_T$, Proposition 2 gives the explicit choice $\alpha = \tau / I_T$. The resulting code satisfies $G(\alpha) \geq \tau$ and has exactly

$$E_{\mathrm{mean}}(\alpha) = \left(\frac{\tau}{I_T}\right)^2 E_{\mathrm{mean}}(1).$$

This is an achievability statement for the analytic ray, not a claim that the displayed energy is globally minimal over arbitrary nonlinear code functions.

B.3.4 AVERAGE-RADIUS VARIANT OF THE CERTIFIED CYLINDER

The main theorem uses a uniform radius for a hard per-sample guarantee. The same proof gives

$$G(u_0 + \delta) \geq G_0 - L_{\mathrm{simp}} \mathbb{E}_X \|P_{\mathcal{V}} \delta(X)\|_2.$$

Thus an average visible-displacement budget also preserves the empirical/population margin in expectation. The uniform version is preferable when the goal is to define a strict feasible set for every sample.

B.4 SCALED ANALYTIC TARGETS AND THE MARGIN-ENERGY TRADEOFF

The closed-form code $u_T(x)$ gives exact matching, $S(u_T(x)) = T_x$. It is often preferable to use a scaled target

$$u(x) = \alpha u_T(x), \qquad 0 \leq \alpha \leq 1,$$

which trades certificate strength against proximity to the Gaussian prior center. Along this ray,

$$S_k(\alpha u_T(x)) = \frac{T_{x,k}^{\alpha} \bar{T}_k^{1-\alpha}}{\sum_j T_{x,j}^{\alpha} \bar{T}_j^{1-\alpha}}.$$

Thus $\alpha = 0$ gives the constant predictor $\bar{T}$, while $\alpha = 1$ gives the teacher T_x. Let

$$Z_\alpha(x) = \sum_j T_{x,j}^{\alpha} \bar{T}_j^{1-\alpha}.$$

Then

$$\mathrm{KL}(T_x \| S(\alpha u_T(x))) = (1-\alpha)\mathrm{KL}(T_x\|\bar{T}) + \log Z_\alpha(x).$$

Hence $L_{\mathrm{TS}}(0) = I_T$ and $L_{\mathrm{TS}}(1) = 0$.
The path is continuous. It is also nonincreasing in α on $[0, 1]$. To see this, write

$$\ell_k(x) = \log \frac{T_{x,k}}{\bar{T}_k}, \qquad A_x(\alpha) = \log \sum_j \bar{T}_j \exp(\alpha \ell_j(x)).$$

Then

$$S_\alpha(k \mid x) = \bar{T}_k \exp\{\alpha \ell_k(x) - A_x(\alpha)\},$$

and the log-partition function satisfies

$$A_x'(\alpha) = \mathbb{E}_{S_\alpha(x)} \ell(x), \qquad A_x''(\alpha) = \mathrm{Var}_{S_\alpha(x)}(\ell(x)) \geq 0.$$

Thus A_x' is nondecreasing. Since $S_1(x) = T_x$, we have $A_x'(1) = \mathbb{E}_{T_x} \ell(x)$, and therefore $A_x'(\alpha) \leq A_x'(1)$ for $0 \leq \alpha \leq 1$. Differentiating the KL expression gives

$$\frac{d}{d\alpha} \mathrm{KL}(T_x \| S(\alpha u_T(x))) = -\mathbb{E}_{T_x} \ell(x) + A_x'(\alpha) \leq 0.$$

In nondegenerate cases the inequality is strict except on flat intervals where the relevant log-odds are constant.

For any margin $0 < \tau < I_T$, at least one α satisfies

$$L_{\mathrm{TS}}(\alpha) \le I_T - \tau.$$

Hence the safety scale $\alpha_\star(\tau)$ defined above is well defined; it is the minimum-energy point on this analytic ray that still achieves the desired certificate margin. Its mean-KL energy is

$$\frac{1}{2}\mathbb{E}_x\|\alpha_\star(\tau)u_T(x)\|^2 = \alpha_\star(\tau)^2\frac{K-1}{2K\beta^2}\mathcal{E}_T.$$

The design interpretation is direct: move only as far from the prior center as needed to beat the constant baseline.

The same interpretation can be used to quantify partial non-collapse inside the protected witness subspace. The scalar α does not create new active latent directions; it only scales the teacher directions already present in $u_T(x)$. Thus, for $\alpha > 0$, the rank of the mean-code covariance along this ray is the same as the rank of $\mathrm{Cov}(u_T(x))$. What improves continuously is not full dimensional usage, but the amount of teacher information expressed by the fixed witness. This distinction is important: $L_{\mathrm{TS}} \approx 0$ means $S(u(x)) \approx T_x$, not that every coordinate or every posterior dimension is noncollapsed.

B.5 Distance to the analytic target and its effect on alignment

Let

$$d(x) = u_\phi(x) - u_T(x).$$

Since $S(u_T(x)) = T_x$, the witness-logit perturbation caused by $d(x)$ is

$$\delta a_k(x) = \beta v_k^\top d(x).$$

Therefore

$$\mathrm{KL}(T_x\|S(u_T(x) + d(x))) = \log\sum_j T_{x,j}e^{\delta a_j(x)} - \sum_k T_{x,k}\delta a_k(x).$$

For small $d(x)$,

$$\mathrm{KL}(T_x\|S(u_T(x) + d(x))) \approx \frac{1}{2}\delta a(x)^\top\left[\mathrm{Diag}(T_x) - T_xT_x^\top\right]\delta a(x).$$

A simple upper bound follows from the bounded curvature of the softmax log-partition:

$$L_{\mathrm{TS}} \le \frac{\beta^2 K}{2(K-1)}\mathbb{E}_x\|u_\phi(x) - u_T(x)\|^2.$$

Thus controlling the witness-relevant distance to $u_T(x)$ controls the alignment loss. The converse is not generally true in the full latent space, since directions orthogonal to the simplex span may not change S.

B.6 General linear-witness safe tubes and trajectory-mined preservation

The fixed simplex gives the explicit constant L_{simp}. A corresponding statement holds for any fixed linear-softmax witness $S(u) = \mathrm{softmax}(Wu + b)$. Row-center the matrix,

$$W_c = \left(I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right)W.$$

Common-logit shifts are irrelevant, and

$$\nabla_u\mathrm{KL}(T_x\|S(u)) = W_c^\top(S(u) - T_x), \qquad \|\nabla_u\mathrm{KL}(T_x\|S(u))\|_2 \le \sqrt{2}\|W_c\|_2 =: K_W.$$

If an anchor μ_0 has safety margin $G_0 = I_T - L_{\mathrm{TS}}(\mu_0) - \tau > 0$, every refinement $\mu = \mu_0 + \delta$ with $\sup_x\|\delta(x)\| \le \rho$ satisfies

$$G_\tau(\mu) \ge G_0 - K_W\rho.$$

For a Gaussian encoder whose variance is held fixed during this refinement, $\|\delta(x)\|_2$ is also the pointwise W_2 distance between the source and refined diagonal posteriors. Thus $\rho < G_0/K_W$ defines a posterior tube that excludes complete input-independent posterior collapse without requiring an escape gradient and without putting the witness loss in the refinement objective.

B.6.1 Validation-only anchor selection and final stress test

One ordinary CIFAR-100 ConvVAE trajectory used latent dimension 64 and was trained from random initialization for 50 epochs with $\beta_{\mathrm{KL}} = 10^{-4}$, followed by strong-KL stress with $\beta_{\mathrm{KL}} = 5$. Candidate

checkpoints were predeclared. On each candidate, the encoder was frozen, the witness was fitted on the training split, and validation-only selection maximized the normalized certified radius subject to a positive margin and a PSNR tolerance. Test data were not used to select the anchor.

The selected epoch was 50. A quick scan gave $G_{0,\mathrm{val}} = 1.90583$, normalized radius 0.0227936, and PSNR 20.5282 dB. After the checkpoint was fixed, the final three-seed witness protocol gave

$$G_{0,\mathrm{val}} = 2.77111, \qquad \rho = 0.0403527,$$

with a held-out finite-split analytical lower bound $1.35285 > 0$. The chosen radius is one half of the final certified radius G_0/K_W. Here $G_{0,\mathrm{val}}$ is the validation safety margin used to set the radius; Table 7 reports held-out test G_τ after the anchor and radius are fixed, so the values need not coincide.

Table 7: Detailed trajectory-mined preservation experiment.

Branch	Final G_τ	Min. G_τ	KL	AU	Max disp.	Stress obj.
Frozen anchor	2.73842	2.73842	54.2750	64	0	271.386
Unconstrained	-0.670815	-0.670815	1.67735×10^{-6}	0	16.0919	0.0700051
Certified tube	2.73512	2.73490	54.0151	64	0.0403527	270.086

The tube had zero constraint violations. Relative to the frozen source it reduced the exact stress objective by 1.29947, with reconstruction-MSE increase 2.95431×10^{-5}, KL reduction 0.2599, and PSNR change -0.011927 dB. The residual was strongly sample dependent:

$$\|\mathbb{E}\delta(X)\|_2 = 0.00310921, \qquad \text{centered residual RMS} = 0.0402327,$$

with centered energy fraction 0.994063, 64 active residual dimensions, and effective rank 59.0465. A common translation would have centered energy fraction zero.

Interpretation. This experiment establishes preservation rather than creation. The early VAE trajectory creates the first informative checkpoint; validation-only mining discovers it; the certificate converts it into a nonzero region that allows further optimization while excluding complete Teacher-relative collapse. The unconstrained branch demonstrates why the constraint is nontrivial: under strong KL pressure it crosses the certificate boundary and becomes completely collapsed under the declared diagnostics.

B.7 Two-view transfer and information–energy–stability Teacher design

A useful extension considers the case in which the Teacher used for training is computed from a different controlled view than the one used for evaluation. Let Q_x be the training-view Teacher and P_x the evaluation-view Teacher, with

$$I_Q = \mathbb{E}\mathrm{KL}(Q_x\|\bar{Q}), \qquad I_P = \mathbb{E}\mathrm{KL}(P_x\|\bar{P}),$$
$$L_Q = \mathbb{E}\mathrm{KL}(Q_x\|S(u(x))), \quad L_P = \mathbb{E}\mathrm{KL}(P_x\|S(u(x))).$$

Then, defining

$$\Delta_{XT} := |I_P - I_Q| + |L_P - L_Q|,$$

the evaluation-view margin obeys the deterministic inequality

$$I_P - L_P \geq I_Q - L_Q - \Delta_{XT}.$$

For the analytic code that exactly matches Q_x, $L_Q = 0$ and $L_P = \mathbb{E}\mathrm{KL}(P_x\|Q_x)$, so the pre-training quantity

$$\Delta_{XT}^{\mathrm{analytic}} = |I_P - I_Q| + \mathbb{E}\mathrm{KL}(P_x\|Q_x)$$

is directly computable. A practical Teacher-search score can therefore balance

$$\text{information } I_Q, \qquad \text{minimum exact energy } \frac{K-1}{2K\beta^2}\mathcal{E}_Q, \qquad \text{view stability } \Delta_{XT}^{\mathrm{analytic}}.$$

This score is a design heuristic rather than a theorem about downstream performance. The main experiments use the cleaner same-view case, where the gap is zero.

B.8 A population concentration bound for a frozen marginal

The main benchmark claims are exact on the empirical evaluation distribution. If a population statement is desired, a simple bound is available under an explicit positivity condition. Assume the population teacher marginal $P_C = \bar{T}$ is fixed and, for some $\eta \in (0, 1)$,

$$\bar{T}_c \geq \eta, \qquad S(c \mid Z) \geq \eta \quad \text{almost surely for every } c.$$

For i.i.d. evaluation pairs (X_i, Z_i) define

$$Y_i = \sum_c T(c \mid X_i) \log \frac{S(c \mid Z_i)}{\bar{T}_c}, \qquad \widetilde{G}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Then $\mathbb{E}Y_i = G_T$ by the Barber–Agakov form and $|Y_i| \le \log(1/\eta)$. Hoeffding's inequality therefore gives, with probability at least $1 - \delta$,

$$G_T \ge \widetilde{G}_n - \log(1/\eta) \sqrt{\frac{2 \log(2/\delta)}{n}}. \tag{A.1}$$

Thus a strictly positive right-hand side certifies positive population teacher information. This bound is deliberately separate from the main empirical-split certificate: it assumes a frozen population marginal rather than replacing it by the same-split empirical $\bar{T}_n$. More refined sample-splitting or empirical-Bernstein analyses are possible but are not needed for the paper's benchmark claims.

B.9 Certificate identities and consistency checks

Direct Barber–Agakov form. Using $I_T = \mathbb{E} \log(T(C \mid X)/P_C(C))$ and the definition of L_{TS} gives directly

$$G_T = \mathbb{E}_{C,Z} \log \frac{S(C \mid Z)}{P_C(C)},$$

which is the variational MI lower bound. Theorem 1 refines it by identifying the exact calibration gap.

Empirical-distribution exactness. For a fixed evaluation set $\{x_i\}_{i=1}^n$, take P_X uniform over those points and define $\bar{T} = n^{-1} \sum_i T_{x_i}$. All main identities then hold exactly on this finite probability space. This distinguishes exact certification on the benchmark split from generalization to an unseen population.

B.9.1 The constant boundary follows directly from the information identity

If $Z \equiv u_0$, then $I(C; Z) = 0$. Theorem 1 therefore gives

$$G_T = -\mathrm{KL}(\bar{T} \| S(\cdot \mid u_0)) \le 0,$$

which is identical to the direct KL decomposition in Theorem 1. The centered witness at $u_0 = 0$ makes the right-hand side zero. Thus the word "exact" can be read either from the best-constant-predictor decomposition or from the information identity.

B.9.2 Why the Teacher information equals the input–Teacher mutual information

By construction $P(C = c \mid X = x) = T_x(c)$ and $P_C(c) = \bar{T}_c$, so

$$I(C; X) = \mathbb{E}_X \sum_c P(c \mid X) \log \frac{P(c \mid X)}{P_C(c)} = \mathbb{E}_X \mathrm{KL}(T_X \| \bar{T}) = I_T.$$

Thus teacher information is ordinary mutual information of the induced random variable C, not an analogy or proxy definition.

B.9.3 Consistency with data processing

Because $C \leftarrow X \rightarrow Z$ and the experimental route is generated from X, data processing gives $I(C; Z) \le I(C; X) = I_T$. Together with the main theorem,

$$G_T \le I(C; Z) \le I_T.$$

Hence a positive population certificate naturally lies in $0 < G_T \le I_T$. A substantial numerical value $G_T > I_T$ should therefore trigger an implementation, normalization, or estimation audit.

B.9.4 First-order alignment versus second-order mean-prior cost

The simplex energy identity gives

$$\frac{1}{2} \mathbb{E} \|u_T(X)\|^2 = \frac{K - 1}{2K\beta^2} \mathcal{E}_T.$$

For $u_\alpha = \alpha u_T$, the standard-Gaussian mean-prior cost is exactly

$$\frac{1}{2} \mathbb{E} \|u_\alpha(X)\|^2 = \alpha^2 \frac{K - 1}{2K\beta^2} \mathcal{E}_T.$$

Meanwhile Proposition 2 gives $L'_{\mathrm{TS}}(0) < 0$ whenever $I_T > 0$. The witness-alignment improvement is therefore first order at the origin while the mean-prior penalty has no first-order term. This is the clean local content of the escape-force statement; the reconstruction term of the full ELBO can still contribute at first order.

C Decodability versus prescribed alignment

Back to main text: Discussion and scope; Experimental evidence.
This block isolates the distinction between information that a newly fitted head can decode and information that is available through the predeclared interface. It places the exact rotation separation next to the matched Aux/RST controls, cross-seed and cross-architecture reuse tests, adapter recovery, and controlled gauge checks.

C.1 Why learned decodability does not imply prescribed alignment

The distinction used in the main text has a simple exact form. A learned probe is allowed to absorb a change of latent coordinates; a prescribed interface is not. The following construction shows that the implication can fail even when Teacher recovery is exact.

Proposition 5 (Decodability–alignment separation)**.** *Let the centered simplex witness be*

$$S_V(u) = \mathrm{softmax}(\beta V u + \log \bar{T}), \qquad \mathcal{V} = \mathrm{row}(V), \quad \mathcal{N}_V = \ker V,$$

and let $u_T(x) \in \mathcal{V}$ be the exact Teacher code, so $S_V(u_T(x)) = T_x$. For every orthogonal matrix Q, the transformed code $\tilde{u}(x) = Q u_T(x)$ is still exactly decodable by the correspondingly transformed head

$$S_Q(\tilde{u}) = \mathrm{softmax}(\beta V Q^\top \tilde{u} + \log \bar{T}), \qquad S_Q(\tilde{u}(x)) = T_x.$$

If $\dim \mathcal{N}_V \geq \dim \mathcal{V} = K - 1$, one can choose Q with $Q\mathcal{V} \subseteq \mathcal{N}_V$. For that choice,

$$S_V(\tilde{u}(x)) = \bar{T}, \qquad L_{\mathrm{TS}}(S_V, \tilde{u}) = I_T, \qquad G_T(S_V, \tilde{u}) = 0,$$

while the rotated learned head still has exact Teacher alignment, $L_{\mathrm{TS}}(S_Q, \tilde{u}) = 0$.

Proof. Orthogonality gives $Q^\top \tilde{u} = u_T$, so the transformed head reproduces the original logits. When $Q\mathcal{V} \subseteq \ker V$, $V\tilde{u} = 0$ and the original fixed witness reduces to $\mathrm{softmax}(\log \bar{T}) = \bar{T}$. The constant-boundary identity then gives $L_{\mathrm{TS}} = I_T$ and $G_T = 0$. □

This is a logical separation, not a claim that unconstrained learned representations typically rotate exactly into $\ker V$. Its role is to identify what a post-hoc probe can hide: coordinate changes can preserve perfect decodability while destroying a previously declared interface. The Aux-full/Aux-block controls in Table 3 test the corresponding practical question without constructing Q by hand.

C.2 Auxiliary-supervision and functional-interface controls

Back to main text: Section 3 (functional-interface evidence)
Table 8 records all ten auxiliary-control runs. The learned-head columns measure the task each control actually optimizes. The fixed-witness columns are adapter-free post-training audits using the same predeclared simplex geometry as RST; no new readout is fitted for these columns. Thus a gap between the two is informative rather than contradictory: it indicates that the labels are readable in the latent code but not in the independently prescribed orientation.

Table 8: Per-seed auxiliary-supervision controls. a^{aux} is accuracy of the learned auxiliary head; a^{pre} and G use the predeclared fixed simplex witness after training, with no refitted adapter.

Method	Seed	a_c^{aux}	a_f^{aux}	G_c	G_f	a_c^{pre}	a_f^{pre}	AU
Aux-full	0	0.4033	0.2737	-3.8858	-4.7483	0.0599	0.0093	192
Aux-full	1	0.4035	0.2765	-2.5292	-4.5347	0.0365	0.0100	192
Aux-full	2	0.3943	0.2566	-3.0721	-4.6013	0.0449	0.0123	192
Aux-full	3	0.4163	0.2915	-3.8300	-5.1843	0.0631	0.0065	192
Aux-full	4	0.4084	0.2692	-2.1256	-5.3920	0.0602	0.0129	192
Aux-block	0	0.3727	0.2202	-2.7727	-4.4938	0.0564	0.0082	184
Aux-block	1	0.3680	0.2308	-1.3214	-4.7288	0.0380	0.0119	186
Aux-block	2	0.3605	0.2380	-3.8074	-4.5395	0.0523	0.0105	187
Aux-block	3	0.3737	0.2244	-3.6090	-5.8882	0.0444	0.0086	190
Aux-block	4	0.3536	0.1805	-1.9743	-6.0291	0.0475	0.0114	187

Table 9: Five-seed summary for the auxiliary controls and the existing joint-RST reference.

Method	G_c	G_f	KL	AU	PSNR
Aux-full	-3.089 ± 0.779	-4.892 ± 0.377	55.74 ± 1.12	192.0 ± 0.0	20.004 ± 0.052
Aux-block	-2.697 ± 1.059	-5.136 ± 0.758	58.32 ± 2.36	186.8 ± 2.2	20.021 ± 0.130
RST-joint	1.057 ± 0.027	1.477 ± 0.026	56.89 ± 0.82	191.8 ± 0.4	20.108 ± 0.102

C.2.1 Functional-interface evaluation: probes, cross-seed reuse, and adapters

All encoders/decoders are frozen for this evaluation. A true post-hoc linear softmax probe is fitted on the CIFAR-100 training split and evaluated on the test split using the same soft Teacher KL and I_T scale as the certificate. Aux-block and RST probes read the declared $19/99$-dimensional blocks; Aux-full reads all 192 mean coordinates. The probe is diagnostic only and never alters the saved representation.

Table 10: Frozen-encoder route-feasibility summary (five seeds). Positive G_{probe} with negative G_{pre} means Teacher information is readable by a newly fitted route but not certified on the predeclared interface.

Method	View	G_{probe}	probe acc. (%)	G_{pre}	predeclared acc. (%)
Aux-block	coarse	0.986 ± 0.013	38.91 ± 0.52	-2.697 ± 1.059	4.77 ± 0.71
Aux-block	fine	1.619 ± 0.037	29.31 ± 0.70	-5.136 ± 0.758	1.01 ± 0.17
Aux-full	coarse	1.078 ± 0.013	43.10 ± 0.33	-3.089 ± 0.779	5.29 ± 1.16
Aux-full	fine	1.710 ± 0.028	32.09 ± 0.54	-4.892 ± 0.377	1.02 ± 0.26
RST-joint	coarse	1.162 ± 0.016	45.35 ± 0.37	1.057 ± 0.027	43.84 ± 1.30
RST-joint	fine	1.957 ± 0.026	37.48 ± 0.54	1.477 ± 0.026	33.58 ± 0.56

For cross-seed reuse, source seed s contributes its jointly trained Aux-block head and target seed t contributes its frozen encoder. Diagonal pairs $(s = t)$ are native; off-diagonal pairs use no adapter. To avoid treating the 20 ordered off-diagonal pairs as independent replicates, we first average the four source heads for each target, then report mean $\pm$ sample s.d. over the five target seeds. Every one of the 40 off-diagonal coarse/fine Teacher margins is negative.

For recovery, paired deterministic train features fit an origin-preserving orthogonal Procrustes map

$$R^\star = \arg\min_{R^\top R = I} \|X_t R - X_s\|_F^2$$

using 20,000 examples (Schönemann, 1966); a ridge-affine map is a more flexible supplementary diagnostic. Procrustes makes $39/40$ transfer margins positive, while ridge-affine makes all $40/40$ positive. Because affine recovery is consistently stronger, the evidence supports a general coordinate/interface mismatch and should not be read as showing that independently trained Aux codes differ only by a rotation.

Table 11: Cross-seed reuse and adapter recovery. Accuracy statistics use the per-target aggregation described above. G columns use the same Teacher-relative margin.

View	Interface	Accuracy (%)	G
coarse	Aux native	36.57 ± 0.85	0.866 ± 0.021
coarse	no adapter	5.20 ± 0.54	-0.967 ± 0.438
coarse	+Procrustes	28.88 ± 3.87	0.540 ± 0.213
coarse	+ridge-affine	35.70 ± 0.43	0.826 ± 0.006
fine	Aux native	21.88 ± 2.24	1.178 ± 0.116
fine	no adapter	0.94 ± 0.10	-1.075 ± 0.245
fine	+Procrustes	16.59 ± 1.04	0.939 ± 0.052
fine	+ridge-affine	21.17 ± 0.29	1.132 ± 0.018

C.2.2 Cross-architecture fixed-interface and reader-transfer audit

The cross-architecture audit keeps the CIFAR-100 Teacher definitions, smoothing $\epsilon = 0.02$, witness gain $\beta = 5$, latent dimension 192, and the $19/99$ coarse/fine visible blocks fixed. Five ResVAE seeds are trained for both RST-joint and Aux-block. For the ResVAE runs, each fine class contributes 450 training and 50 validation examples (a deterministic $45{,}000/5{,}000$ split with split seed 20260905); the official 10,000-image test set is not accessed until the minimum predeclared validation objective has locked the checkpoint. Training uses AdamW, 55 epochs, five-epoch KL warmup, and cosine learning rate $2 \times 10^{-4} \to 10^{-5}$. RST uses only reconstruction, KL, and the two fixed-witness alignment losses with $\lambda_c = \lambda_f = 2$: no analytic-code regression and no hard parameterization $u = \alpha u_T + Nh$ are used. The ConvVAE checkpoints were trained on the full 50k training set and

are reused unchanged; therefore this audit tests whether a second architecture can satisfy the same fixed interface, not a matched-data Conv-versus-Res performance ranking.

Table 12: Cross-architecture fixed-interface summary (mean ± sample s.d. over five seeds). Native Aux columns use each control's jointly trained head; fixed columns use the common predeclared simplex witness.

Producer	Method	G_c	G_f	fixed a_c (%)	fixed a_f (%)	native a_c (%)	native a_f (%)
ConvVAE	RST	1.057 ± 0.027	1.477 ± 0.026	43.84 ± 1.30	33.58 ± 0.56	–	–
ResVAE	RST	0.869 ± 0.057	1.083 ± 0.105	37.28 ± 1.94	23.87 ± 2.20	–	–
ConvVAE	Aux-block	-2.697 ± 1.059	-5.136 ± 0.758	4.77 ± 0.71	1.01 ± 0.17	36.57 ± 0.85	21.88 ± 2.24
ResVAE	Aux-block	-2.543 ± 1.004	-5.314 ± 0.883	4.90 ± 0.75	0.85 ± 0.18	20.03 ± 1.11	10.38 ± 1.09

For learned-reader transfer, every ordered source-head/target-encoder combination is evaluated (5×5 pairs in each architecture direction and view). To avoid treating the five source heads for one target as independent replicates, Table 13 first averages source heads within each target and then reports mean ± sample s.d. over the five target seeds. Both adapters use 20,000 paired feature examples drawn only from the common training subset; neither validation nor test examples are used for adapter fitting.

Table 13: Cross-architecture learned-reader reuse and recovery. Each accuracy/G pair is target-aggregated over five seeds as described in the text.

Transfer	View	no-adapter acc.	no-adapter G	Proc. acc.	Proc. G	affine acc.	affine G
Conv→Res	coarse	5.11 ± 0.50	-0.993 ± 0.404	17.22 ± 2.41	0.017 ± 0.282	21.81 ± 1.27	0.458 ± 0.033
Conv→Res	fine	0.91 ± 0.14	-0.816 ± 0.138	9.92 ± 0.57	0.489 ± 0.045	14.74 ± 0.64	0.777 ± 0.036
Res→Conv	coarse	4.98 ± 0.50	-0.120 ± 0.069	21.69 ± 2.90	0.323 ± 0.021	24.43 ± 0.40	0.317 ± 0.003
Res→Conv	fine	0.96 ± 0.04	-0.125 ± 0.028	10.33 ± 0.86	0.528 ± 0.052	10.47 ± 0.15	0.501 ± 0.004

At the raw pair level, $98/100$ no-adapter cross-architecture margins are negative; the two positive values are only 0.0046 and 0.0232. Orthogonal Procrustes makes $89/100$ positive, while ridge-affine makes all $100/100$ positive. The asymmetric and incomplete Procrustes recovery is evidence for a broader coordinate/interface mismatch, not for a pure rotation. Conversely, positive fixed-witness margins for all ten Conv/Res RST runs show that two independently optimized producer architectures can satisfy the same predeclared interface; this is evidence for cross-architecture realizability of the specified interface, not a universal architecture-invariance theorem.

C.2.3 Controlled gauge and selective block interventions

For a controlled coordinate-gauge sanity check, rotate an RST visible block as $X' = XQ$. Keeping the original fixed witness makes accuracy collapse to 4.81% (coarse) and 1.00% (fine), whereas transforming the readout as $V' = VQ$ restores the original 43.84% and 33.58% accuracies. The adapted logits match the originals to at most 4.8×10^{-6} across the five seeds. This is a controlled algebraic check of the separation proposition, not a claim that natural SGD produces an exact orthogonal rotation.

Zero/swap interventions audit the block contract. Zeroing the coarse block drives its margin to 5.8×10^{-8} while leaving the fine margin at 1.477 ± 0.026; zeroing the fine block drives its margin to -4.6×10^{-7} while leaving the coarse margin at 1.057 ± 0.027. Swapping a target block with a donor preserves donor-interface accuracy ($43.84 \pm 1.30\%$ coarse; $33.58 \pm 0.56\%$ fine), destroys agreement with the recipient on that route ($4.84 \pm 0.19\%$; $0.88 \pm 0.04\%$), and leaves the orthogonal non-target margin numerically unchanged. These are latent-interface interventions, not decoder-causality evidence.

D Full empirical evidence and reproducibility

Back to main text: Section 3 (Experimental evidence)

This block contains the complete tables, protocols, and implementation details underlying the compact main-text results. It is intended as the first stop for numerical verification of reported means, seed stability, matched controls, and reproducibility.

D.1 CIFAR-100 routing, anti-collapse, and efficiency studies

Selective routing and composition. The CIFAR-100 coarse labels define $K_c = 20$ and the fine labels define $K_f = 100$. Their regular-simplex blocks occupy 19 and 99 mutually orthogonal coordinates of a $d_z = 192$ code, leaving 74 common witness-null coordinates. The Teacher smoothing parameter is 0.02, witness gain is 5, batch size is 256, and the stress stage uses $\beta_{\mathrm{KL}} = 5$. The five

variants are baseline, coarse-only, fine-only, joint, and joint with explicit null routing. Every reported main-table value is mean $\pm$ sample standard deviation over seeds 0–4.

Matched anti-collapse comparison. The matched comparison uses the same convolutional backbone, latent dimension 192, batch size 256, 55 epochs, evaluation every five epochs, AdamW learning rate 2×10^{-4}, weight decay 10^{-4}, KL weight 1 with five-epoch warmup, and automatic mixed precision (AMP). InfoVAE uses a multi-scale inverse-multiquadratic maximum mean discrepancy (IMQ-MMD) estimator with $\alpha = 0$, $\lambda = 1000$, and scales $\{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. An MMD gradient sanity test verified that optimization reduced a large translated-distribution MMD to near zero and that $\mathrm{MMD}(x, x) = 0$ numerically. Hyperparameters for the reported InfoVAE were fixed from its own optimization stability rather than from RST certificate margins. The δ control fixes posterior variance so that the zero-mean conditional KL equals 8 nats; nonzero means can only increase that KL. It is an architecture-matched committed-rate control, not an exact reproduction of the original autoregressive δ-VAE architecture. RST-joint uses coarse and fine alignment weights of 2.

Auxiliary-supervision control. We add two learned-head controls to separate the supervision signal from the prescribed simplex route. Both use the same ConvVAE, latent dimension 192, smoothed coarse/fine Teachers, 55 epochs, batch size 256, five seeds, KL schedule, and weights $\lambda_c = \lambda_f = 2$ as the joint RST comparison. Instead of a fixed simplex witness, the auxiliary loss is computed through trainable linear softmax heads. Aux-full reads all 192 latent-mean coordinates. Aux-block restricts the coarse head to the 19-dimensional coarse block and the fine head to the 99-dimensional fine block, matching the RST block allocation. The frozen simplex witnesses are excluded from control training and are used only after training for independent certificate and fixed-route-accuracy audits. We interpret high learned-head accuracy with nonpositive fixed-route margins as decodability without certified realization on the prescribed route; the audit margin is not used for hyperparameter selection.

Efficiency environment. The 20 anti-collapse runs report system metadata from the run records. The profiling environment exposed one NVIDIA GeForce RTX 4080 vGPU with 31.473 GiB CUDA-visible memory. Driver version was 580.105.08; the host utility reported CUDA 13.0 while the PyTorch binary was built against CUDA 12.8. Python was 3.12.3, PyTorch 2.8.0+cu128, torchvision 0.23.0+cu128, and cuDNN 9.10.2. The host had Intel Xeon Platinum 8260 CPUs and 503 GiB system RAM. Peak memory in Table 19 is `torch.cuda.max_memory_allocated`, not total driver-level GPU memory usage.

Interpretive discipline. A positive G_c or G_f is sufficient evidence for the corresponding Teacher-relative information route. A non-positive value is not evidence of literal mutual-information zero. Active units are reported as a generic latent-use diagnostic and are not equated with the certificate. High common-null mean energy is a geometric routing statistic only; without an explicit decoder-use intervention it is not assigned a semantic interpretation.

D.1.1 Complete five-seed statistics for the CIFAR-100 studies

Tables 14–20 report the full five-seed summaries used to construct the main-text tables. Wide tables are split by inferential role so certificate/interface quantities remain readable without extreme scaling. These values are not additional hyperparameter selections; they are measurements from the same frozen five-seed runs.

Table 14: Five-seed selective routing/composition: certificate and interface diagnostics. "Null frac." is the posterior-mean energy fraction in the common witness-null complement; it is geometric and is not a decoder-use claim.

Method	G_c	G_f	Acc. c	Acc. f	Null frac.
VAE	-1.660 ± 0.964	-1.728 ± 0.324	0.0545 ± 0.0053	0.0085 ± 0.0018	0.420 ± 0.078
RST-coarse	$+0.619 \pm 0.017$	-2.519 ± 0.416	0.3454 ± 0.0091	0.0088 ± 0.0018	0.458 ± 0.081
RST-fine	-2.302 ± 1.167	$+0.579 \pm 0.013$	0.0536 ± 0.0026	0.2099 ± 0.0062	0.729 ± 0.135
RST-joint	$+0.686 \pm 0.019$	$+0.655 \pm 0.019$	0.3694 ± 0.0083	0.2254 ± 0.0093	0.948 ± 0.004
RST-joint-null	$+0.693 \pm 0.017$	$+0.632 \pm 0.014$	0.3770 ± 0.0080	0.2359 ± 0.0114	0.954 ± 0.005

Table 15: Five-seed selective routing/composition: reconstruction and generic latent-use diagnostics. Missing MSE entries were not archived for the routed variants.

Method	Rec.	MSE	KL	PSNR	AU
VAE	1749.526 ± 1.968	0.018024 ± 0.000250	20.614 ± 0.318	17.442 ± 0.060	36.2 ± 2.3
RST-coarse	1748.231 ± 2.853	–	21.249 ± 0.846	17.488 ± 0.101	49.2 ± 4.1
RST-fine	1749.124 ± 1.439	–	20.819 ± 0.375	17.457 ± 0.043	39.4 ± 1.5
RST-joint	1746.727 ± 2.322	–	22.029 ± 0.907	17.536 ± 0.077	57.8 ± 4.8
RST-joint-null	1748.554 ± 2.366	–	21.483 ± 0.514	17.484 ± 0.069	52.4 ± 6.6

Table 16: Matched anti-collapse comparison: prescribed-view certificates, fixed-interface accuracy, and activity.

Method	G_c	G_f	Acc. c	Acc. f	AU
VAE	-3.344 ± 1.131	-4.162 ± 0.396	0.0519 ± 0.0022	0.0084 ± 0.0010	167.8 ± 2.8
InfoVAE	-3.435 ± 0.913	-4.650 ± 0.340	0.0479 ± 0.0025	0.0105 ± 0.0013	172.2 ± 10.3
δ-control	-4.718 ± 0.368	-5.944 ± 0.193	0.0506 ± 0.0024	0.0094 ± 0.0013	192.0 ± 0.0
RST-joint	$+1.057 \pm 0.027$	$+1.477 \pm 0.026$	0.4385 ± 0.0130	0.3358 ± 0.0056	191.8 ± 0.4

Table 17: Matched anti-collapse comparison: prior diagnostics. These columns are diagnostic and are not themselves certificates.

Method	KL	min-KL	MMD
VAE	55.155 ± 0.866	27.552 ± 0.628	0.0590 ± 0.0040
InfoVAE	54.027 ± 0.375	28.061 ± 0.965	0.0733 ± 0.0075
δ-control	67.734 ± 2.326	31.190 ± 1.287	0.1276 ± 0.0025
RST-joint	56.885 ± 0.817	28.250 ± 1.190	0.0621 ± 0.0033

Table 18: Matched anti-collapse comparison: reconstruction and null-routing diagnostics.

Method	PSNR	MSE	Null frac.
VAE	20.022 ± 0.042	0.009951 ± 0.000096	0.434 ± 0.039
InfoVAE	19.891 ± 0.072	0.010256 ± 0.000171	0.385 ± 0.056
δ-control	19.597 ± 0.105	0.010975 ± 0.000267	0.385 ± 0.007
RST-joint	20.108 ± 0.102	0.009756 ± 0.000231	0.926 ± 0.001

Table 19: Complete five-seed reconstruction and efficiency summary for the matched anti-collapse comparison. Peak values are PyTorch allocator measurements.

Method	Rec.	Runtime (s)	Throughput (img/s)	Peak alloc. (GiB)	Peak reserved (GiB)	Params
VAE	1686.795 ± 0.811	98.43 ± 2.32	29226 ± 707	0.2086 ± 0.0002	0.2813 ± 0.0000	3,682,435
InfoVAE	1689.348 ± 1.360	137.03 ± 3.08	20715 ± 464	0.2095 ± 0.0002	0.2656 ± 0.0000	3,682,435
δ-control	1696.112 ± 2.148	97.91 ± 2.26	29410 ± 699	0.1981 ± 0.0010	0.2645 ± 0.0070	3,682,435
RST-joint	1685.473 ± 1.965	101.25 ± 3.73	28371 ± 1086	0.2090 ± 0.0002	0.2656 ± 0.0000	3,682,435

Table 20: Per-seed certificate audit for the matched anti-collapse comparison. The $(-,-)$ versus $(+,+)$ distinction is stable in every seed.

Method	seed	G_c	G_f	AU	KL	PSNR
VAE	0	-4.5005	-4.1361	168	55.065	20.0184
VAE	1	-2.6751	-4.0955	165	54.750	20.0238
VAE	2	-3.7656	-3.6560	165	54.175	20.0427
VAE	3	-4.0600	-4.7680	171	55.266	20.0681
VAE	4	-1.7178	-4.1520	170	56.518	19.9555
InfoVAE	0	-3.8727	-4.3208	174	53.749	19.8620
InfoVAE	1	-1.9744	-4.4195	162	54.168	19.8630
InfoVAE	2	-3.1825	-4.6296	173	53.770	19.9777
InfoVAE	3	-4.3325	-5.1959	188	54.626	19.9508
InfoVAE	4	-3.8128	-4.6851	164	53.824	19.8000
δ	0	-4.6535	-6.0753	192	67.295	19.7004
δ	1	-4.6218	-6.0239	192	68.301	19.6351
δ	2	-4.9686	-5.6838	192	66.323	19.5386
δ	3	-4.1934	-5.8007	192	65.349	19.4435
δ	4	-5.1543	-6.1369	192	71.401	19.6680
RST-joint	0	+1.0341	+1.4668	192	57.526	20.0921
RST-joint	1	+1.0238	+1.4417	192	55.547	19.9384
RST-joint	2	+1.0652	+1.4813	192	57.260	20.1821
RST-joint	3	+1.0858	+1.5122	192	56.679	20.1809
RST-joint	4	+1.0758	+1.4825	191	57.415	20.1486

Table 21: Evidence beyond five-dataset boundary crossing. AU denotes active units.

Test	Setting	Result
Strong-KL robustness	$\beta_{\mathrm{KL}} = 4$ on CIFAR-10/100	VAE $-0.637/-0.547$ versus RST-α-logit-prefit $+1.317/+0.853$
Witness-null routing	routed CIFAR-10/100, 5 seeds	$G = 0.276/0.102$, logit residual 0.001, reconstruction close to VAE; decoder use is not inferred
Standard anti-collapse	warm-up, cyclical annealing, free bits	all remain certificate-negative on CIFAR-10/100 ($G \approx -1.63$ to -2.26)
Teacher correspondence	degraded, shuffled, random	margin decreases with degradation; random Dirichlet Teachers are near zero $(-0.004/-0.005)$
Route re-instantiation	moving-shape blocks; log-variance statistic	VAE $G = (-3.707, -2.422)$ versus Block-RST $(0.909, 0.468)$; variance check $G = 0, 0.240, 2.728, 2.394$
Optimization escape	CIFAR-10/100/Tiny-ImageNet	final $G \approx 2.0$, 1.6–1.7, 1.31–1.43
Preservation	CIFAR-100 strong-KL refinement	unconstrained $G_\tau = -0.671$, AU$= 0$; certified refinement has min. $G_\tau = 2.73490$, AU$= 64$

D.1.2 Additional breadth evidence

D.2 Experimental protocol and reproducibility details

This section records the implementation details used by the reported experiments. The theorem-relevant distinction is architectural rather than cosmetic: reconstruction uses a sampled VAE latent, whereas certification reads only the designated code/statistic through a fixed code-only witness.

D.2.1 Architecture and protected routes

All image experiments use the standard-Gaussian VAE interface

$$q_\phi(z \mid x) = \mathcal{N}\big(u_\phi(x), \operatorname{diag} \sigma_\phi^2(x)\big), \qquad p(z) = \mathcal{N}(0, I),$$

The five-dataset PCA–GMM benchmark and its associated stress/counterfactual runs use a convolutional CIFAR-scale encoder/decoder with $d_z = 128$. The CIFAR-100 coarse/fine routing, matched-control, auxiliary, and cross-architecture studies use $d_z = 192$; the trajectory-mined preservation study in Appendix B.6 uses $d_z = 64$. The reconstruction branch uses samples $z \sim q_\phi(z \mid x)$; the certificate branch is deterministic and evaluates the posterior mean $u_\phi(x)$ unless a route-local re-instantiation is stated explicitly. Tiny-ImageNet-200 is resized to 32×32 and uses the same CIFAR-scale VAE interface. The theorem-relevant components are summarized below.

Component	Map	Role in the reported runs
Encoder	$x \mapsto (u_\phi(x), \log \sigma_\phi^2(x))$	Produces the sampled VAE posterior and the deterministic protected mean-code route.
Decoder	$z \mapsto \hat{x}$	Standard reconstruction path; it is not an input to the certificate.
Teacher	$x \mapsto T_x$	Smoothed PCA–GMM responsibilities in the image experiments; selected before certified training and then frozen.
Simplex witness	$u \mapsto \operatorname{softmax}(\beta V u + \log \bar{T})$	Fixed code-only Student; $\beta = 5.0$ in the main image runs.
witness-null route	$u = \alpha u_T + N h_\phi, VN = 0$	Keeps the witness-visible component controlled while allowing witness-invisible residual information.

The layer definitions and run configurations used for the reported studies are summarized in this appendix; the certificate itself depends only on the protected-route separation above, not on a particular convolutional backbone.

D.2.2 Teacher construction and search

The image Teachers are constructed from PCA features followed by a diagonal-covariance GMM. Responsibilities are smoothed to full support, candidate Teachers are evaluated before VAE training, and the selected Teacher is then frozen for every certified run using that Teacher. The search is designed to trade off an informative but not excessively sharp view. The candidate score uses the ingredients

$$I_T(\omega), \qquad \frac{K-1}{2K\beta^2}\mathcal{E}_T(\omega), \qquad \mathrm{Bal}(\bar{T}_\omega), \qquad \mathrm{Fit}(\omega), \qquad \mathrm{Instab}(\omega),$$

corresponding respectively to Teacher information, analytic mean-prior energy, component balance, GMM fit, and instability across seeds or mild perturbations. A representative scalarization is

$$\mathrm{Score}(\omega) = \lambda_I I_T(\omega) - \lambda_E \frac{K-1}{2K\beta^2}\mathcal{E}_T(\omega) + \lambda_B \mathrm{Bal}(\bar{T}_\omega) + \lambda_F \mathrm{Fit}(\omega) - \lambda_S \mathrm{Instab}(\omega),$$

with the concrete search weights treated as implementation choices rather than theorem assumptions. PCA–GMM is only a reproducible way to instantiate a fixed nontrivial Teacher; the certificate theorem requires the final Teacher/route pair, not this search procedure.

The frozen Teachers used in the five-dataset benchmark have the following theorem-facing statistics:

Dataset	K	I_T	$\mathcal{E}_T$
CIFAR-10	50	2.7275	53.5416
CIFAR-100	50	2.7349	57.9823
Fashion-MNIST	10	1.8727	24.8682
SVHN	50	2.7538	53.0375
Tiny-ImageNet-200	50	2.6080	62.1318

For Tiny-ImageNet-200, the frozen final Teacher uses 64 PCA dimensions, a diagonal GMM, temperature 0.7, probability smoothing $\epsilon = 0.1$, and GMM seed 1. The remaining per-dataset search-grid metadata are preserved in the frozen Teacher artifacts/configuration package; these search choices are implementation details rather than theorem assumptions.

D.2.3 Training and reporting protocol

Unless a control explicitly changes one of these quantities, the main five-dataset image runs use 80 training epochs, $d_z = 128$, witness gain $\beta = 5.0$, $\beta_{\mathrm{KL}} = 1.0$, and $\lambda_{\mathrm{TS}} = 1.0$. The CIFAR-100 joint/matched-control, auxiliary, and cross-architecture studies are explicit exceptions and use the per-view weights $\lambda_c = \lambda_f = 2$ stated above. Analytic-prefit variants use five prefit epochs. The KL-pressure stress test sets $\beta_{\mathrm{KL}} = 4.0$. Main five-dataset summaries use five random seeds; the KL-pressure and counterfactual-control blocks use three seeds. Multi-seed tables report the final training epoch and sample standard deviation. For the PCA–GMM broad benchmark tables and their associated stress/counterfactual blocks, $\bar{T}$, I_T, and L_{TS} are computed self-consistently on the same empirical training split on which the frozen Teacher responsibilities are stored. Those G_T values are therefore exact certificates for that empirical training distribution, not claims of held-out population generalization. The CIFAR-100 coarse/fine cross-architecture audit instead follows the predeclared 45k/5k train/validation split and locked-checkpoint official-test protocol in Appendix C.2.2. Reconstruction, KL, active units, effective rank, MMD/prior checks where available, and visible-path residuals are diagnostics; $G_T = I_T - L_{\mathrm{TS}}$ remains the certificate quantity.

D.2.4 Gradient-dominance diagnostics

To distinguish the function-space escape field from its realized parameter-space update, define

$$g_{\mathrm{TS}} = \nabla_\phi L_{\mathrm{TS}}, \qquad g_{\mathrm{VAE}} = \nabla_\phi(\mathcal{L}_{\mathrm{rec}} + \beta_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}),$$

and monitor

$$\Gamma_{\mathrm{grad}} = \frac{\lambda_{\mathrm{TS}}\|g_{\mathrm{TS}}\|}{\|g_{\mathrm{VAE}}\| + 10^{-12}}, \qquad \cos(g_{\mathrm{TS}}, g_{\mathrm{VAE}}).$$

The ratio measures relative scale, while the cosine diagnoses agreement or opposition between the two parameter-space gradients. Early CIFAR-10 runs give mean Γ_{grad} values of about 0.30–0.44 across seeds, with many collapse-neighborhood steps above one. The cosine is used as a conflict diagnostic: a strongly negative value warns that the VAE objective opposes the witness update, while a non-opposing value indicates that the two signals are not cancelling by direction. These measurements support the local optimization interpretation only; they are neither part of the certificate nor an optimizer-independent convergence theorem.

D.3 Complete diagnostic tables

The main text uses slim summaries so that the certificate message is readable. This appendix reports the diagnostic columns omitted from the slim main-text summaries for CIFAR, five-dataset, KL-pressure, and counterfactual results. These diagnostics are not themselves the certificate. They are included to make the experimental record complete and auditable.

Table 22: Complete diagnostics for the main CIFAR-10 and CIFAR-100 runs.

Dataset	Method	Seeds	$G_T \uparrow$	$L_{\mathrm{TS}} \downarrow$	Rec↓	KL	Active units	Eff. rank	Path residual↓
CIFAR-10	VAE	5	-1.987 ± 0.360	4.714 ± 0.360	50.547 ± 0.675	22.922 ± 0.911	50.8 ± 6.6	33.19 ± 0.31	205.974 ± 39.474
CIFAR-10	RST	5	1.885 ± 0.033	0.842 ± 0.033	50.710 ± 0.450	23.048 ± 0.406	94.4 ± 3.6	38.06 ± 0.14	45.505 ± 9.327
CIFAR-10	RST-prefit	5	1.914 ± 0.029	0.814 ± 0.029	50.794 ± 0.313	22.818 ± 0.299	90.6 ± 3.6	38.58 ± 0.15	68.555 ± 9.314
CIFAR-10	RST-alpha-prefit	5	1.841 ± 0.062	0.886 ± 0.062	50.416 ± 0.495	23.342 ± 0.480	95.0 ± 4.2	38.15 ± 0.25	65.776 ± 29.899
CIFAR-10	RST-alpha-logit-prefit	5	1.836 ± 0.041	0.891 ± 0.041	50.511 ± 0.514	23.324 ± 0.506	91.4 ± 2.3	38.33 ± 0.21	75.930 ± 19.893
CIFAR-10	RST-nullspace-route	5	0.276 ± 0.000	2.452 ± 0.000	50.191 ± 0.294	23.043 ± 0.288	44.6 ± 3.3	33.06 ± 0.18	0.001 ± 0.000
CIFAR-100	VAE	5	-2.074 ± 0.242	4.809 ± 0.242	50.080 ± 0.507	23.202 ± 0.497	50.6 ± 3.8	32.16 ± 0.25	215.797 ± 48.189
CIFAR-100	RST	5	1.556 ± 0.021	1.179 ± 0.021	49.892 ± 0.248	23.784 ± 0.255	92.8 ± 4.0	38.16 ± 0.11	42.849 ± 8.226
CIFAR-100	RST-prefit	5	1.692 ± 0.038	1.043 ± 0.038	50.242 ± 0.736	23.448 ± 0.396	89.8 ± 3.3	38.42 ± 0.28	36.692 ± 10.508
CIFAR-100	RST-alpha-prefit	5	1.488 ± 0.056	1.247 ± 0.056	50.013 ± 0.319	23.564 ± 0.268	100.4 ± 4.3	37.74 ± 0.23	37.024 ± 9.471
CIFAR-100	RST-alpha-logit-prefit	5	1.542 ± 0.039	1.193 ± 0.039	50.056 ± 0.412	23.547 ± 0.520	94.2 ± 2.6	38.02 ± 0.28	36.171 ± 6.450
CIFAR-100	RST-nullspace-route	5	0.102 ± 0.000	2.633 ± 0.000	49.768 ± 0.450	23.361 ± 0.368	49.0 ± 2.3	31.94 ± 0.24	0.001 ± 0.000

Table 23: Complete diagnostics for the five-dataset extension corresponding to Table 2.

Dataset	Method	Seeds	$G_T \uparrow$	$L_{\mathrm{TS}} \downarrow$	Rec↓	KL	Active units	Eff. rank	Path residual↓
CIFAR-10	VAE	5	-1.987 ± 0.360	4.714 ± 0.360	50.547 ± 0.675	22.922 ± 0.911	50.800 ± 6.573	33.186 ± 0.307	205.974 ± 39.474
CIFAR-10	RST-alpha-prefit	5	1.841 ± 0.062	0.886 ± 0.062	50.416 ± 0.495	23.342 ± 0.480	95.000 ± 4.183	38.148 ± 0.254	65.776 ± 29.899
CIFAR-10	RST-alpha-logit-prefit	5	1.836 ± 0.041	0.891 ± 0.041	50.511 ± 0.514	23.324 ± 0.506	91.400 ± 2.302	38.328 ± 0.211	75.930 ± 19.893
CIFAR-10	RST-nullspace-route	5	0.276 ± 0.000	2.452 ± 0.000	50.191 ± 0.294	23.043 ± 0.288	44.600 ± 3.286	33.060 ± 0.175	0.001 ± 0.000
CIFAR-100	VAE	5	-2.074 ± 0.242	4.809 ± 0.242	50.080 ± 0.507	23.202 ± 0.497	50.600 ± 3.847	32.162 ± 0.253	215.797 ± 48.189
CIFAR-100	RST-alpha-prefit	5	1.488 ± 0.056	1.247 ± 0.056	50.013 ± 0.319	23.564 ± 0.268	100.400 ± 4.336	37.742 ± 0.229	37.024 ± 9.471
CIFAR-100	RST-alpha-logit-prefit	5	1.542 ± 0.039	1.193 ± 0.039	50.056 ± 0.412	23.547 ± 0.520	94.200 ± 2.588	38.021 ± 0.277	36.171 ± 6.450
CIFAR-100	RST-nullspace-route	5	0.102 ± 0.000	2.633 ± 0.000	49.768 ± 0.450	23.361 ± 0.368	49.000 ± 2.345	31.938 ± 0.243	0.001 ± 0.000
Fashion-MNIST	VAE	5	-0.277 ± 0.328	2.150 ± 0.328	220.245 ± 0.190	15.546 ± 0.131	15.000 ± 2.000	12.577 ± 0.301	6.018 ± 7.715
Fashion-MNIST	RST-alpha-prefit	5	1.503 ± 0.060	0.370 ± 0.060	220.833 ± 0.391	15.879 ± 0.215	19.800 ± 0.447	13.958 ± 0.251	100.903 ± 20.845
Fashion-MNIST	RST-alpha-logit-prefit	5	1.626 ± 0.054	0.247 ± 0.054	220.474 ± 0.167	15.830 ± 0.264	22.200 ± 2.168	14.411 ± 0.199	46.225 ± 24.271
Fashion-MNIST	RST-nullspace-route	5	0.103 ± 0.000	1.770 ± 0.000	220.466 ± 0.267	15.440 ± 0.232	13.600 ± 0.894	12.589 ± 0.037	0.000 ± 0.000
SVHN	VAE	5	-1.088 ± 0.133	3.842 ± 0.133	22.501 ± 0.388	11.598 ± 0.264	23.000 ± 2.449	18.198 ± 0.173	109.389 ± 19.873
SVHN	RST-alpha-prefit	5	1.943 ± 0.026	0.811 ± 0.026	22.326 ± 0.360	12.219 ± 0.313	68.600 ± 2.702	24.928 ± 0.120	67.584 ± 6.953
SVHN	RST-alpha-logit-prefit	5	1.967 ± 0.029	0.786 ± 0.029	22.292 ± 0.225	12.347 ± 0.120	68.200 ± 1.483	25.262 ± 0.321	69.967 ± 7.972
SVHN	RST-nullspace-route	5	0.106 ± 0.000	2.648 ± 0.000	22.189 ± 0.300	11.904 ± 0.302	23.400 ± 1.517	17.984 ± 0.119	0.002 ± 0.000
Tiny-ImageNet-200	VAE	5	-2.279 ± 0.360	4.887 ± 0.360	50.341 ± 0.445	23.911 ± 0.505	50.600 ± 2.408	35.213 ± 0.264	245.728 ± 45.396
Tiny-ImageNet-200	RST-alpha-prefit	5	1.321 ± 0.036	1.287 ± 0.036	50.318 ± 0.507	24.274 ± 0.454	100.000 ± 3.000	41.309 ± 0.530	33.359 ± 1.538
Tiny-ImageNet-200	RST-alpha-logit-prefit	5	1.317 ± 0.101	1.291 ± 0.101	50.406 ± 0.346	24.159 ± 0.375	98.400 ± 1.517	41.736 ± 0.280	42.920 ± 14.856
Tiny-ImageNet-200	RST-nullspace-route	5	0.102 ± 0.000	2.506 ± 0.000	50.113 ± 0.498	24.059 ± 0.508	46.000 ± 1.871	35.215 ± 0.086	0.001 ± 0.000

Table 24: Complete diagnostics for the KL-pressure stress test with $\beta_{\mathrm{KL}} = 4.0$.

Dataset	Method	Seeds	$G_T \uparrow$	$L_{\mathrm{TS}} \downarrow$	Rec↓	KL	Active units	Eff. rank	Path residual↓
CIFAR-10	VAE	3	-0.637 ± 0.274	3.364 ± 0.274	82.668 ± 0.801	6.980 ± 0.201	13.333 ± 1.528	11.092 ± 0.077	53.699 ± 22.029
CIFAR-10	RST-alpha-prefit	3	1.243 ± 0.024	1.484 ± 0.024	82.174 ± 1.005	7.190 ± 0.254	53.667 ± 1.528	14.183 ± 0.223	38.268 ± 9.246
CIFAR-10	RST-alpha-logit-prefit	3	1.317 ± 0.004	1.410 ± 0.004	81.714 ± 0.954	7.311 ± 0.212	55.000 ± 1.000	14.162 ± 0.099	33.969 ± 0.879
CIFAR-10	RST-nullspace-route	3	0.159 ± 0.100	2.568 ± 0.100	82.352 ± 0.492	7.008 ± 0.096	11.000 ± 0.000	11.109 ± 0.067	0.000 ± 0.000
CIFAR-100	VAE	3	-0.547 ± 0.051	3.282 ± 0.051	81.899 ± 0.319	7.580 ± 0.105	11.000 ± 1.000	11.137 ± 0.027	51.976 ± 9.911
CIFAR-100	RST-alpha-prefit	3	0.770 ± 0.108	1.965 ± 0.108	81.867 ± 1.265	7.686 ± 0.287	36.333 ± 1.528	13.326 ± 0.210	24.504 ± 19.293
CIFAR-100	RST-alpha-logit-prefit	3	0.853 ± 0.013	1.882 ± 0.013	82.298 ± 0.503	7.545 ± 0.089	37.000 ± 6.928	13.233 ± 0.129	13.219 ± 0.644
CIFAR-100	RST-nullspace-route	3	0.101 ± 0.000	2.634 ± 0.000	82.246 ± 0.526	7.516 ± 0.125	14.000 ± 2.646	11.000 ± 0.115	0.000 ± 0.000

Table 25: Complete diagnostics for baseline and teacher counterfactual controls.

Dataset	Block	Condition	Seeds	$G_T \uparrow$	$L_{\mathrm{TS}} \downarrow$	Rec↓	KL	Active units	Eff. rank
CIFAR-10	baseline	VAE+KL-warmup	3	-1.628 ± 0.096	4.355 ± 0.096	50.135 ± 0.534	23.053 ± 0.492	42.333 ± 0.577	34.733 ± 0.458
CIFAR-10	baseline	VAE+cyclical-annealing	3	-1.962 ± 0.137	4.690 ± 0.137	49.909 ± 0.402	23.622 ± 0.240	50.667 ± 1.155	37.456 ± 0.440
CIFAR-10	baseline	VAE+free-bits	3	-2.256 ± 0.260	4.983 ± 0.260	53.151 ± 1.001	22.205 ± 0.752	128.000 ± 0.000	36.338 ± 0.890
CIFAR-10	teacher	degraded_rho0.5	3	0.540 ± 0.005	0.384 ± 0.005	50.154 ± 0.225	23.012 ± 0.244	50.000 ± 3.000	35.067 ± 0.044
CIFAR-10	teacher	degraded_rho0.75	3	0.147 ± 0.001	0.177 ± 0.001	50.158 ± 0.638	23.024 ± 0.612	38.667 ± 2.517	34.016 ± 0.285
CIFAR-10	teacher	random_dirichlet	3	-0.004 ± 0.001	0.372 ± 0.001	50.238 ± 0.423	23.137 ± 0.509	45.000 ± 4.359	33.112 ± 0.075
CIFAR-10	teacher	shuffled	3	0.159 ± 0.022	2.569 ± 0.022	50.131 ± 0.167	23.200 ± 0.260	42.000 ± 1.732	34.256 ± 0.273
CIFAR-100	baseline	VAE+KL-warmup	3	-2.227 ± 0.483	4.961 ± 0.483	49.255 ± 0.039	23.800 ± 0.289	41.667 ± 2.082	33.868 ± 0.184
CIFAR-100	baseline	VAE+cyclical-annealing	3	-2.218 ± 0.273	4.953 ± 0.273	49.235 ± 0.799	23.985 ± 0.806	51.667 ± 0.577	36.900 ± 0.240
CIFAR-100	baseline	VAE+free-bits	3	-2.174 ± 0.099	4.909 ± 0.099	53.343 ± 0.268	22.127 ± 0.291	128.000 ± 0.000	34.406 ± 0.279
CIFAR-100	teacher	degraded_rho0.5	3	0.409 ± 0.004	0.513 ± 0.004	50.230 ± 0.171	22.987 ± 0.115	51.000 ± 5.000	34.190 ± 0.131
CIFAR-100	teacher	degraded_rho0.75	3	0.102 ± 0.004	0.221 ± 0.004	49.971 ± 0.502	23.304 ± 0.353	47.333 ± 3.215	32.846 ± 0.206
CIFAR-100	teacher	random_dirichlet	3	-0.005 ± 0.000	0.373 ± 0.000	50.568 ± 0.528	22.813 ± 0.443	50.000 ± 4.583	31.547 ± 0.414
CIFAR-100	teacher	shuffled	3	0.138 ± 0.011	2.597 ± 0.011	50.007 ± 0.368	23.395 ± 0.588	53.000 ± 4.359	32.840 ± 0.218

E MECHANISTIC AND NUMERICAL AUDITS

Back to main text: Section 2 (theory and optimization diagnostics)

This block collects implementation-facing checks that connect the analytic objects to optimization and diagnostics: gradient flow, the closed-form path audit, code sanity, failed-certificate diagnosis, decoder-side diagnostics, escape, and witness-weight behavior. These audits support interpretation; they do not enlarge the theorem claims.

E.1 GRADIENT-FLOW SCHEMATIC

Figure 4 separates forward computation from optimization signals. The Teacher–Student loss compares the frozen Teacher target with the fixed code-only witness and sends its gradient only through the encoder-side mean-code route. Reconstruction trains the encoder and decoder, while KL regularizes the posterior parameters. This figure is explanatory rather than an additional empirical claim; the parameter-space caveat and gradient diagnostics are stated in the main text and later in the appendix.

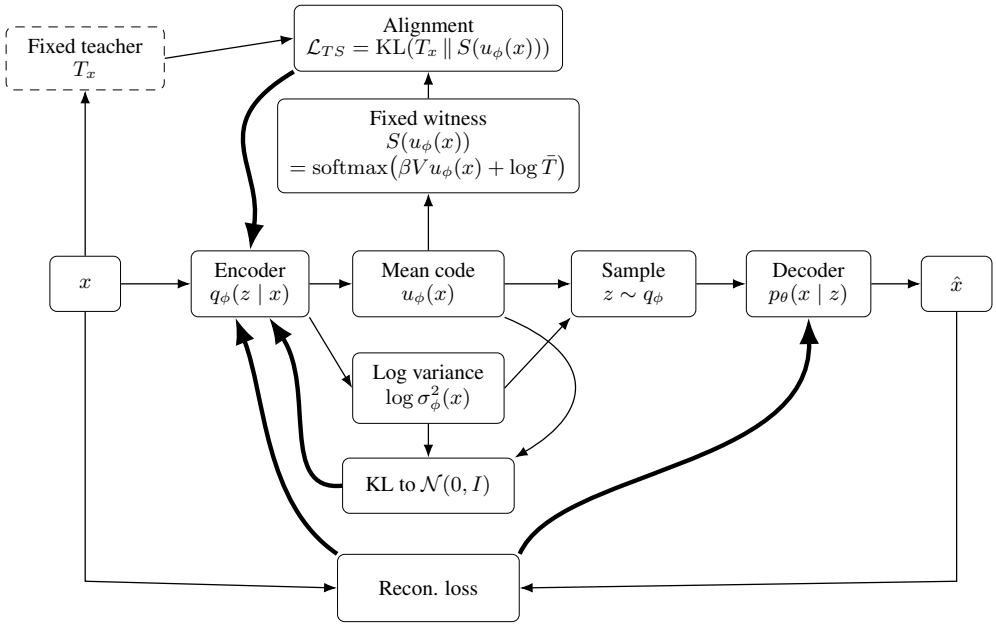

Figure 4: Gradient-flow view of RST. Thin arrows show forward computation; thick arrows show optimization signals. The fixed Teacher and witness prevent the certified-stage alignment loss from being reduced by Teacher–Student co-adaptation.

E.2 ANALYTIC MARGIN–ENERGY PATH AUDIT: FULL CIFAR-100 VALUES

Back to main text: Section 2 and Figure 2

This appendix records every sampled point used in Figure 2. It is a deterministic evaluation of the closed-form simplex construction on the 10,000-image CIFAR-100 test split, with $\epsilon = 0.02$, $\beta = 5$, and $d_z = 192$. "Slack" denotes $G(\alpha) - \alpha I_T$. The -1.49×10^{-8} coarse endpoint slack is floating-point roundoff at a theoretical equality.

Table 26: Coarse Teacher ($K = 20$): $I_T = 2.8456666$, $E(1) = 0.8565164$.

α	$G(\alpha)$	αI_T	slack	$E(\alpha)$
0.00	-0.0000000	0.0000000	-0.0000000	0.0000000
0.05	0.3175338	0.1422833	0.1752505	0.0021413
0.10	0.6273866	0.2845667	0.3428199	0.0085652
0.20	1.2132628	0.5691333	0.6441295	0.0342607
0.40	2.1512680	1.1382666	1.0130013	0.1370426
0.60	2.6512941	1.7074000	0.9438941	0.3083459
0.80	2.8170229	2.2765333	0.5404896	0.5481705
1.00	2.8456666	2.8456666	-0.0000000	0.8565164

The numerical audits are: inverse maximum error 6.966×10^{-9} (coarse) and 7.688×10^{-9} (fine); maximum geometric-path error 9.992×10^{-16} and 1.332×10^{-15}; maximum quadratic-energy-

Table 27: Fine Teacher ($K = 100$): $I_T = 4.4169271$, $E(1) = 1.4153098$.

α	$G(\alpha)$	αI_T	slack	$E(\alpha)$
0.00	0.0000000	0.0000000	0.0000000	0.0000000
0.05	0.4111677	0.2208464	0.1903214	0.0035383
0.10	0.8195940	0.4416927	0.3779013	0.0141531
0.20	1.6220519	0.8833854	0.7386665	0.0566124
0.40	3.0774768	1.7667708	1.3107059	0.2264496
0.60	4.0313523	2.6501562	1.3811960	0.5095115
0.80	4.3657539	3.5335417	0.8322122	0.9057983
1.00	4.4169271	4.4169271	0.0000000	1.4153098

law error 1.110×10^{-16} and 0; minimum lower-bound slack -1.490×10^{-8} and 7.843×10^{-9}. These numbers audit implementation of exact identities and should not be interpreted as statistical confidence intervals or learned-model generalization results.

E.3 Teacher-code encoder sanity check

The teacher-code interpretation can be checked without a full VAE by training a small encoder to predict the analytic teacher code from vector inputs. This diagnostic is not used as a main empirical claim; it records whether the map $x \mapsto u_T(x)$ is learnable only when the teacher is aligned with the input.

In a synthetic vector replay setting, we constructed a five-component risk teacher from an input feature and trained an encoder to predict the corresponding teacher code. On a held-out validation split, the aligned teacher achieved

$$I_T = 1.429147, \qquad L_{\mathrm{TS}} = 0.153129, \qquad G_T = 1.276019.$$

We then replaced the teacher by two negative controls. A random-label teacher gave

$$I_T = 1.599293, \qquad L_{\mathrm{TS}} = 1.690099, \qquad G_T = -0.090807,$$

and a shuffled version of the original risk teacher gave

$$I_T = 1.429147, \qquad L_{\mathrm{TS}} = 1.536108, \qquad G_T = -0.106961.$$

Thus the positive held-out margin appears for the input-aligned teacher but not for random or sample-shuffled teachers. This supports the intended interpretation of teacher-code learning: the encoder can learn a readable teacher-code representation when the teacher view is a genuine function of the input, while arbitrary or misaligned pseudo-teachers fail the held-out certificate.

E.4 Non-positive margins, route feasibility, and decoder-side diagnostics

Back to main text: Figure 1 (route and decoder diagnostics)

The boundary case

$$L_{\mathrm{TS}} \geq I_T, \qquad u_\phi(x) \not\equiv \text{constant}$$

is nonconstant but not Teacher-certified on the prescribed route. It may encode other information, or Teacher information in coordinates unreadable by the fixed witness.

E.4.1 Three-point geometry of the origin, an intermediate code, and the Teacher code

Let

$$S(u) = \mathrm{softmax}(\log \bar{T} + \beta V u), \qquad W = \beta V,$$

and let $u_T(x)$ be an analytic teacher deterministic code satisfying $S(u_T(x)) = T_x$. The three relevant points are the constant baseline 0, the current deterministic code $u(x) = u_\phi(x)$, and the teacher-perfect deterministic code $u_T(x)$. With

$$a_u(x) = \log \bar{T} + W u(x), \qquad a_T(x) = \log \bar{T} + W u_T(x),$$

and $A(a) = \log \sum_k e^{a_k}$, the log-sum-exp Bregman divergence is

$$D_A(a_u, a_T) = A(a_u) - A(a_T) - \nabla A(a_T)^\top (a_u - a_T).$$

Since $\nabla A(a_T) = T_x$, we have the exact identity

$$D_A(a_u, a_T) = \mathrm{KL}(T_x \| S(u(x))).$$

Therefore

$$\mathbb{E}_x D_A(a_u, a_T) = L_{\mathrm{TS}}(u), \qquad \mathbb{E}_x D_A(a_0, a_T) = I_T,$$

where $a_0 = \log \bar{T}$. Consequently,

$$L_{\mathrm{TS}}(u) > I_T \quad \Longleftrightarrow \quad \mathbb{E}_x D_A(a_u, a_T) > \mathbb{E}_x D_A(a_0, a_T).$$

Thus the unfavorable regime means that u is, on average, farther from the teacher deterministic code than the constant baseline in the witness-Bregman geometry.
There is also a useful three-point decomposition:

$$L_{\mathrm{TS}}(u) - I_T = \underbrace{\mathbb{E}_x \mathrm{KL}(\bar{T} \| S(u(x)))}_{\text{departure cost from 0}} - \underbrace{\mathbb{E}_x \langle T_x - \bar{T}, W u(x) \rangle}_{\text{teacher-alignment gain}}.$$

This identity explains why $u_\phi \neq \text{constant}$ is not enough. The code may leave the origin, but if its teacher-alignment gain is smaller than the cost of leaving the constant baseline, then $L_{\mathrm{TS}} \geq I_T$.
If one replaces KL by Jensen–Shannon divergence, then the symmetric distance $d_{\mathrm{JS}}(P, Q) = \sqrt{\mathrm{JS}(P, Q)}$ gives an ordinary metric geometry among T_x, $\bar{T}$, and $S(u)$. It can be used as a diagnostic: one asks whether $d_{\mathrm{JS}}(T_x, S(u)) < d_{\mathrm{JS}}(T_x, \bar{T})$. However, the KL/log-sum-exp Bregman form is more useful for the present theory because it gives the exact alignment-gain decomposition and the escape gradient.

E.4.2 Route-feasibility diagnostics

Two diagnostics separate a failed prescribed certificate from a broader representation failure. First, freeze the encoder and fit a flexible probe $P_\psi(c \mid u_\phi(x))$ with

$$L_{\mathrm{probe}} = \mathbb{E}_x \mathrm{KL}(T_x \| P_\psi(\cdot \mid u_\phi(x))).$$

If $L_{\mathrm{probe}} < I_T$ while $G_T \leq 0$, the code contains Teacher-view information that is readable by another route, so the failure is specific to the fixed witness. Second, freeze the Teacher, witness geometry, and decoder and optimize only the encoder-side witness route,

$$\min_\phi \; L_{\mathrm{TS}} = \min_\phi \; \mathbb{E}_x \mathrm{KL}(T_x \| S(u_\phi(x))).$$

If this crosses I_T, the route is representationally feasible and the original failure is more plausibly due to objective conflict or insufficient alignment. The frozen-probe experiments in Appendix C.2.1 instantiate the first case directly: both Aux controls have $G_{\mathrm{probe}} > 0$ while $G_{\mathrm{pre}} < 0$ on both views, and cross-seed and cross-architecture adaptation tests further localize the mismatch to producer–consumer coordinates. If neither diagnostic beats I_T, one should inspect Teacher sharpness/stability, witness gain, latent or block dimension, and encoder capacity. Thus $G_T \leq 0$ is best treated as a failed sufficient certificate plus a route-feasibility diagnostic, not as a collapse verdict.

E.4.3 Safe Teacher paths and the role of the path scale

For $0 \leq \alpha \leq 1$, the path

$$u_\alpha(x) = \alpha u_T(x)$$

moves from the constant baseline to the teacher-perfect deterministic code. Along this path,

$$L_{\mathrm{TS}}(0) = I_T, \qquad L_{\mathrm{TS}}(u_T) = 0,$$

and convexity in logits gives

$$L_{\mathrm{TS}}(\alpha u_T) \leq (1 - \alpha) I_T.$$

Thus any sufficiently positive α gives a positive certificate margin. For a target margin $0 < \tau < I_T$, one can choose the fixed safety scale

$$\alpha_\star(\tau) = \inf\{\alpha \in [0, 1] : \; L_{\mathrm{TS}}(\alpha u_T) \leq I_T - \tau\},$$

subject also to a prior-cost budget such as

$$\frac{1}{2} \alpha^2 \mathbb{E}_x \|u_T(x)\|^2 \leq \kappa.$$

This makes α a fixed analytic design parameter or a prefit scale, not a free parameter that is minimized jointly with the VAE objective.
If the trained code is decomposed as

$$u_\phi(x) = \alpha u_T(x) + r(x),$$

the relevant off-path residual is not $\|r(x)\|$ but its witness-logit projection

$$e(x) = W(u_\phi(x) - \alpha u_T(x)).$$

The quantity

$$E_e = \mathbb{E}_x \|e(x)\|^2$$

is the teacher-path deviation that the witness can see. If E_e is small relative to the margin already gained by αu_T, then $L_{\mathrm{TS}} < I_T$ is preserved. In particular, a sufficient tube condition is

$$Q_\alpha \sqrt{E_e} + \frac{1}{4} E_e < \Delta_\alpha,$$

where

$$\Delta_\alpha = I_T - L_{\mathrm{TS}}(\alpha u_T), \qquad Q_\alpha = \sqrt{\mathbb{E}_x \|T_x - S(\alpha u_T(x))\|^2}.$$

This relation gives a non-vacuous training target: keep the witness-visible deviation from the teacher path below the safety tube determined by Δ_α.

E.4.4 A STRUCTURAL TEACHER-PATH GUARANTEE WITH A CERTIFICATE-INVARIANT RESIDUAL ROUTE

A direct structural way to keep the witness loss independent of residual code variation is to separate the Teacher-visible route from a witness-null residual route. Let N be a basis for the null space of $W = \beta V$, so

$$WN = 0.$$

Define the deterministic encoder code as

$$u_\phi(x) = \alpha u_T(x) + N h_\phi(x),$$

where $h_\phi(x)$ is an arbitrary learnable encoder function. Then

$$W u_\phi(x) = \alpha W u_T(x),$$

and hence

$$S(u_\phi(x)) = S(\alpha u_T(x)).$$

Therefore, if α is chosen so that $L_{\mathrm{TS}}(\alpha u_T) < I_T$, then the teacher-witness certificate is guaranteed regardless of the behavior of $h_\phi(x)$. The role of $h_\phi(x)$ is to provide a witness-invisible route that remains available to the decoder for factors not represented by the Teacher view. Certificate invariance alone does not establish which factors the decoder actually places there.

This construction is useful when the latent dimension exceeds the witness rank. If the witness uses K simplex components, the teacher-visible row space has dimension at most $K-1$. When $d_z > K-1$, the null space can carry additional code variation without changing the witness output. Choosing $u_T(x)$ as the minimum-norm teacher deterministic code in the row space of W makes $u_T(x)$ orthogonal to $N h_\phi(x)$, so the mean KL cost decomposes into a Teacher-visible cost and a null-route cost; neither term is free with respect to the VAE prior.

E.4.5 S–R ROUTE AND DECODER USAGE

The teacher-path guarantee above certifies a $T \to S$ route: the teacher view enters the latent witness. It does not by itself prove that the decoder uses that route. A powerful decoder might reconstruct primarily through $N h_\phi(x)$ and ignore the protected teacher block. To test decoder usage, one can add an $S \to R$ or $T \to R$ reconstruction-side diagnostic, or a swap-based intervention.

Let $\hat{x} = D_\theta(z)$, with $z \sim q_\phi(z \mid x)$, and define the reconstruction-side posterior

$$R_{\hat{x}} = T_{\omega^\star}(c \mid \hat{x}),$$

obtained by applying the frozen Teacher to the reconstruction. The two optional direct diagnostics are

$$L_{\mathrm{SR}} = \mathbb{E}_{x,z} \mathrm{KL}\left(S(u_\phi(x)) \| R_{D_\theta(z)}\right), \qquad z \sim q_\phi(z \mid x),$$

and

$$L_{\mathrm{TR}} = \mathbb{E}_{x,z} \mathrm{KL}\left(T_x \| R_{D_\theta(z)}\right), \qquad z \sim q_\phi(z \mid x).$$

Here L_{TR} tests whether the reconstruction preserves the Teacher view, whereas L_{SR} tests consistency between the Student witness and the reconstruction-side posterior. These quantities are diagnostics and do not enter the direct certificate $G_T = I_T - L_{\mathrm{TS}}$.

A stronger intervention-based diagnostic is the cross-swap route. For two samples x_i, x_j, decode

$$\hat{x}_{i \leftarrow j} = D_\theta(\alpha u_T(x_j) + N h_\phi(x_i)).$$

If the decoder uses the Teacher-visible route, the reconstruction-side posterior should follow the swapped Teacher block:

$$R_{\hat{x}_{i \leftarrow j}} \approx T_{x_j}.$$

This gives the intervention loss

$$L_{\text{TR-swap}} = \mathbb{E}_{i,j} \mathrm{KL}\left(T_{x_j} \| R_{D_\theta(\alpha u_T(x_j) + N h_\phi(x_i))}\right).$$

Thus $T \to S$ certifies that the Teacher view enters the latent route, while $S \to R$, $T \to R$, and swap-based diagnostics test whether Reconstruction preserves and uses that route. This distinction is essential for decoder bypass and decoder ignorance.

E.4.6 PRACTICAL USE

The simplest procedure chooses α from the analytic Teacher path, optionally prefits the encoder toward $\alpha u_T(x)$ in witness-logit space, and then trains $\mathcal{L}_{\mathrm{VAE}} + \lambda_{\mathrm{TS}} L_{\mathrm{TS}}$. A stronger routed extension uses $u_\phi(x) = \alpha u_T(x) + N h_\phi(x)$ and adds $S \to R$, $T \to R$, or swap diagnostics only when decoder use must be tested. Adaptive weights/scales are possible but are not part of the main empirical method.

E.5 DERIVATION OF THE ESCAPE FIELD AT CONSTANT COLLAPSE

Back to main text: Section 2 (departure/escape result)

For one sample, let

$$a_k(x) = \beta v_k^\top u(x) + \log \bar{T}_k, \qquad S(x) = \mathrm{softmax}(a(x)).$$

Because

$$\mathrm{KL}(T_x \| S(x)) = \mathrm{const} - T_x^\top a(x) + \log \sum_j e^{a_j(x)},$$

we have the standard softmax-gradient identity

$$\nabla_{a(x)} \mathrm{KL}(T_x \| S(x)) = S(x) - T_x,$$

and therefore

$$\nabla_u \ell_x = \beta \sum_{k=1}^{K} \big(S_k(x) - T_{x,k}\big) v_k, \qquad -\nabla_u \ell_x = \beta \sum_{k=1}^{K} \big(T_{x,k} - S_k(x)\big) v_k.$$

At an arbitrary constant deterministic code $u(x) \equiv u_0$, write $s_0 = S(u_0)$. The negative functional gradient is

$$F_T^{u_0}(x) = \beta \sum_{k=1}^{K} \big(T_{x,k} - s_{0,k}\big) v_k.$$

If $F_T^{u_0}(x) = 0$ almost everywhere, then affine independence of the simplex vertices, together with $\sum_k (T_{x,k} - s_{0,k}) = 0$, implies $T_x = s_0$ almost everywhere. Averaging gives $s_0 = \bar{T}$, hence $I_T = \mathbb{E}\mathrm{KL}(T_X \| \bar{T}) = 0$. Consequently,

$$I_T > 0 \quad \Longrightarrow \quad F_T^{u_0} \not\equiv 0$$

for every exact constant code u_0: no such input-independent constant mean is a function-space stationary point of the alignment loss.

At the centered origin, $S(0) = \bar{T}$, so

$$F_T(x) = \beta \sum_{k=1}^{K} (T_{x,k} - \bar{T}_k) v_k.$$

The field is sample-dependent even though $\mathbb{E}_X F_T(X) = 0$. Thus a common translation is not the escape direction; a perturbation $u_\varepsilon(x) = u_0 + \varepsilon F_T^{u_0}(x)$ has first-order derivative

$$\frac{d}{d\varepsilon} L_{\mathrm{TS}}(u_\varepsilon)\Big|_{\varepsilon=0} = -\mathbb{E}\|F_T^{u_0}(X)\|_2^2 < 0$$

whenever $I_T > 0$. This is the function-space departure statement used in the main text; parameter-space realization still depends on the encoder Jacobian and is checked empirically rather than asserted as an optimizer-independent theorem.

E.6 GRADIENT DOMINANCE AND PRACTICAL VARIANTS

At the prior-centered collapsed origin $u = 0$, the standard-Gaussian mean-KL gradient with respect to the deterministic code is zero. If the reconstruction gradient transverse to the collapse manifold has norm at most B, and the RST escape field has norm at least κ, then choosing $\lambda_{\mathrm{TS}} \kappa > B$ makes the RST transverse component dominant. In practice, B and κ can be estimated with batch gradient norms, and possible cancellation can be monitored with gradient angles.

Analytic prefit can train the deterministic encoder code toward $u_T(x)$ or toward the scaled target $\alpha_\star(\tau) u_T(x)$ before full VAE training. A margin-constrained variant can also be used in implementation,

$$[L_{\mathrm{TS}} - (I_T - \tau)]_+^2 .$$

This implementation device can maintain a chosen margin, while the certificate theorem itself depends only on $G_T = I_T - L_{\mathrm{TS}}$.

E.7 Witness weights and escape training

With several protected views, use

$$\mathcal{L} = \mathcal{L}_{\mathrm{VAE}} + \sum_{a \in \mathcal{A}} \lambda_a \mathbb{E}_x \mathrm{KL}(T_x^a \| S_a(H_a(x))) \,.$$

Separate λ_a values accommodate different witness scales or selectively protect mean, variance, block, hierarchy, decoder, or temporal routes. Multiple witnesses can supply additional transverse directions at a collapsed/low-margin state, but only when their gradients are not cancelled by the VAE objective or by each other; this is not an optimizer-independent escape guarantee.

In practice, fixed weights, warm-up/prefit, gradient-norm balancing, and cosine-angle diagnostics are sufficient controls. A constrained variant can treat $G_a \geq \tau_a$ as a margin constraint and adapt λ_a as a Lagrange multiplier, but the reported experiments deliberately use fixed weights and final margins so that the empirical claim remains about the prescribed code-route certificate.

F Extensions, limitations, and re-instantiations

Back to main text: Section 5 (scope and limitations)

This final block separates theorem-level re-instantiations from prospective uses. Each extension inherits only the route-local certificate under its stated assumptions; empirical status and out-of-scope claims are made explicit.

F.1 Scope and evidence level of the re-instantiation principle

Corollary 1 is a statement about a *chosen view/representation pair*, not a claim that one witness solves every phenomenon called "collapse." For a representation/statistic Z_m that satisfies the stated Markov condition, the theorem supplies an exact input-independence boundary and a sufficient information margin. What it does not supply automatically is an appropriate Teacher for every failure mode, a causal decoder-use test, an aggregate-distribution objective, or empirical validation of every possible route. Table 28 separates the mathematical interface from the evidence actually provided in this paper.

Table 28: Scope of the re-instantiation principle. "Theorem gives" refers to the route-local certificate under its assumptions; the last column records the level of evidence in this paper.

Route / view	What the theorem gives	Evidence in this paper
Deterministic latent mean / code	Exact input-independence boundary, information certificate, and—with the simplex witness—closed-form realization, routing, path/cost, escape, preservation, and composition.	Main end-to-end setting; multi-dataset checks and five-seed CIFAR-100 studies.
Log-variance statistic $\log \sigma_\phi^2(x)$	Its own route-specific boundary and margin for a fixed uncertainty-related Teacher.	CIFAR-10 cross-statistic sanity check only; not a general variance-collapse study.
Prescribed latent block	A block-local boundary and certificate when the witness reads only that block.	Five-seed coarse/fine block-routing and composition study.
Grouped Teacher view	Data-processing relation $I_A \le I_T$ plus a group-level route certificate; sufficiently strong fine-view alignment can imply a coarse-view certificate.	Coarse/fine CIFAR-100 instance plus formal derivation.
Multi-level hierarchy	Repeated/grouped certificates and chain-rule structure for explicitly defined hierarchy levels.	Formal consequence; the CIFAR-100 coarse/fine study validates the first nontrivial two-level instance, while higher-depth hierarchies are not separately studied.
Decoder-side statistic	A route-local certificate can be attached to a fixed decoder-side statistic generated from X.	Diagnostic interface only; no intervention study establishing causal decoder use.
Temporal / world-model state	The same certificate can be instantiated for a designated state statistic if the conditional-independence assumptions are appropriate.	Prospective interface only; no rollout, dynamics, causal, or long-horizon validation.
Aggregate posterior / generation coverage	Not covered by Corollary 1 as a simple sample-wise route without defining a different aggregate-level object.	Outside the present theorem and experiments.
Blur / perceptual generation quality	No direct consequence of the route-local information certificate.	Outside scope; requires likelihood, decoder, perceptual, or task-specific analysis.

The certificate is route-local, while the constructive simplex geometry is developed most fully for the deterministic mean route. Table 28 records the evidence level for every extension; mode collapse, aggregate-posterior mismatch, causal decoder use, and world-model dynamics remain outside the present claims.

F.2 Posterior-statistic witnesses for variance collapse

For a diagonal Gaussian encoder,

$$q_\phi(z \mid x) = \mathcal{N}\big(u_\phi(x), \mathrm{diag}(\sigma_\phi^2(x))\big)\,,$$

the posterior mean and log-variance are two statistics of the same latent random variable. The main certificate acts on $u_\phi(x)$. A constant variance by itself is therefore not a collapse of the learned deterministic representation; full exact posterior collapse would require the relevant posterior statistics to become input-independent together.

The route-local principle can nevertheless be re-instantiated on the uncertainty statistic $\ell_\phi(x) = \log \sigma_\phi^2(x)$. Given a fixed full-support uncertainty Teacher T_x^σ with marginal $\bar{T}^\sigma$, define

$$S_\sigma(\ell) = \mathrm{softmax}(\beta_\sigma V_\sigma \ell + \log \bar{T}^\sigma),$$

$$I_T^\sigma = \mathbb{E}_x \mathrm{KL}(T_x^\sigma \| \bar{T}^\sigma), \qquad L_{T\sigma} = \mathbb{E}_x \mathrm{KL}(T_x^\sigma \| S_\sigma(\ell_\phi(x)))\,, \qquad G_T^\sigma = I_T^\sigma - L_{T\sigma}.$$

By Corollary 1, $G_T^\sigma > 0$ certifies that the observed log-variance route is not input-independent with respect to that uncertainty view; when the witness has a constant code producing $\bar{T}^\sigma$, I_T^σ is again the attainable input-independent boundary. The uncertainty Teacher should encode an uncertainty-related view (for example ambiguity, perturbation instability, occlusion, noise level, or prediction uncertainty), rather than being mechanically identified with the semantic code Teacher.

Cross-statistic sanity check. A small CIFAR-10 re-instantiation checks this extension. A collapsed log-variance statistic gives $G_T^\sigma = 0$; a weak input-dependent statistic gives 0.240; the analytic

statistic gives 2.728; and a noisy analytic statistic gives 2.394. This is a route-local sanity check of the same exact boundary, not evidence that the main VAE experiments solve general posterior-variance collapse.

Joint posterior statistics and scope. One may likewise define a joint route $h_\phi(x) = [u_\phi(x); \rho\ell_\phi(x)]$ and attach a route-specific witness. A positive joint margin only certifies that the joint observed statistic is not input-independent (an OR-type statement); it does not imply that both channels are individually non-collapsed. Channel-wise claims require separate margins, e.g. $G_T^\mu > 0$ and $G_T^\sigma > 0$. The main experiments therefore keep the primary certificate on the deterministic mean code and use the variance construction only as a targeted re-instantiation of Corollary 1.

F.3 Structured and temporal re-instantiations

The route-local theorem applies whenever a fixed Teacher view is paired with a designated statistic or route generated from X. Latent-block certificates are the direct structured case: write $H(x) = (H_1(x), \dots, H_B(x))$ and assign a fixed Teacher $T_x^{(b)}$ and witness S_b to block H_b. Then $G_b > 0$ certifies prescribed Teacher information on that block, which is more specific than an active-unit count but remains limited to the assigned witness view.

Decoder use is a separate question; Appendix E.4 defines the $S \to R$, $T \to R$, and cross-swap diagnostics.

For temporal/world-model states, one may choose Z_m to be a designated state block at time t and define fixed Teachers for object, motion, occupancy, interaction, or risk views. The same I_m, L_m, G_m boundary then tests whether the selected state representation is input-independent with respect to that view. This is an audit statement only: rollout quality, interventions, causal dynamics, and long-horizon consistency require separate experiments.

Progressive Teacher extraction (prospective). Suppose m Teacher views have already been frozen and separately certified. A new residual or coarse-to-fine view may be fitted and then frozen as $T^{(m+1)}(c \mid x)$. If it is full-support and nontrivial, its own $(I_{m+1}, L_{m+1}, G_{m+1})$ inherits the same exact constant-boundary proof. Progressive extraction can therefore be organized as a sequence of separately auditable certificates. What is not proved is automatic semantic discovery: defining a meaningful residual, avoiding redundant Teachers, and establishing stability remain separate problems.

F.4 Distributions of the sampled latent, aggregate posterior, and Teacher code

The VAE encoder specifies a conditional Gaussian posterior,

$$q_\phi(z \mid x) = \mathcal{N}(u_\phi(x), \operatorname{diag} \sigma_\phi^2(x)).$$

Thus $z \mid x$ is Gaussian. The aggregate posterior,

$$q_\phi(z) = \int q_\phi(z \mid x) p_{\text{data}}(x)\, dx,$$

is generally a mixture of conditional Gaussians and need not equal a single standard Gaussian. This is compatible with the standard VAE setup: the prior is $p(z) = \mathcal{N}(0, I)$, and the ELBO penalizes each conditional posterior through $\mathrm{KL}(q_\phi(z \mid x) \| p(z))$.

The analytic target $u_T(x)$ is the pushforward of the data distribution through the centered log-odds map. It is generally not Gaussian. For label-smoothed teachers it may take finitely many values; for soft GMM teachers it varies continuously with x. The certificate is evaluated through the deterministic path $u_\phi(x) \mapsto S(u_\phi(x))$, because this path is stable and directly tests input dependence of the deterministic encoder code. Sampling quality and aggregate prior matching should be reported separately when unconditional generation is discussed.

F.5 Group and hierarchical teacher views

The main theorem applies to a single K-component teacher. The same constant-baseline argument can be applied to coarser or hierarchical views derived from the same teacher.

Let $A \in \{0,1\}^{K_A \times K}$ be a fixed grouping matrix whose rows define a partition of the K components, so each original component belongs to exactly one group. Define

$$T_x^A = AT_x, \qquad \bar{T}^A = A\bar{T}.$$

The information content of this group view is

$$I_A = \mathbb{E}_x \mathrm{KL}(T_x^A \| \bar{T}^A).$$

By the data-processing inequality for KL divergence,

$$0 \le I_A \le I_T.$$

If $I_A = 0$, the group view is constant and gives no group-level certificate. If $I_A > 0$, the same simplex construction with K_A vertices gives a group-level witness S^A and a group-level alignment loss

$$L_A = \mathbb{E}_x \mathrm{KL}(T_x^A \| S^A(z_A(x))).$$

Then

$$L_A < I_A$$

rules out input-independent constant collapse of the group-level code $z_A(x)$.

There are two ways to use this observation. First, one may define a separate latent subspace z_A and a separate group witness. At the collapsed point $z_A(x) \equiv 0$, the group-level escape force is

$$F_A(x) = \beta_A \sum_{b=1}^{K_A} (T_{x,b}^A - \bar{T}_b^A) v_b^A,$$

which is not identically zero whenever $I_A > 0$. This is the direct group analogue of Proposition 2.

Second, if the group prediction is obtained by merging the original student probabilities, $S^A = AS$, then KL contraction gives

$$L_A = \mathbb{E}_x \mathrm{KL}(AT_x \| AS(u_\phi(x))) \le \mathbb{E}_x \mathrm{KL}(T_x \| S(u_\phi(x))) = L_{\mathrm{TS}}.$$

Thus, a sufficiently strong original margin automatically certifies group views:

$$L_{\mathrm{TS}} < I_A \quad \Longrightarrow \quad L_A < I_A.$$

Along the analytic α-path, a sufficient group-crossing threshold is

$$\alpha_A^\star = \inf\{\alpha \in [0,1] : L_{\mathrm{TS}}(\alpha) < I_A\}.$$

This sufficient threshold need not equal the first α for which $L_A(\alpha) < I_A$. As α increases, progressively more group views may cross their own constant baselines. This gives a quantitative but limited sense in which partial collapse is improved: the protected teacher information is expressed at more coarse or hierarchical levels. It does not imply that unprotected latent coordinates outside the witness subspace are used.

A hierarchical teacher is obtained by representing the component index as a sequence of decisions,

$$C = (B_1, B_2, \ldots, B_m).$$

For a binary hierarchy, m can be of order $\lceil \log_2 K \rceil$. The information decomposes by the chain rule,

$$I(X;C) = I(X;B_1) + I(X;B_2 \mid B_1) + \cdots + I(X;B_m \mid B_{<m}).$$

This gives a structured interpretation of non-collapse: early levels certify coarse teacher information, while later conditional levels certify finer distinctions inside previously selected groups. In this view, partial non-collapse is not a coordinate-wise statement; it is a statement about which teacher views, from coarse to fine, are carried by the latent representation.

F.6 Fixed GMM responsibility witnesses

The simplex witness can be interpreted as a spherical-GMM responsibility head over the latent space. Let

$$m_k = a v_k, \qquad \Sigma_k = \sigma^2 I, \qquad \pi_k = \bar{T}_k.$$

Then

$$\frac{\pi_k \mathcal{N}(z; m_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(z; m_j, \Sigma_j)} = \mathrm{softmax}_k \left(\frac{a}{\sigma^2} v_k^\top z + \log \bar{T}_k \right).$$

Thus the simplex witness is the most symmetric fixed GMM-responsibility witness, with $\beta = a/\sigma^2$. The latent prior is still $\mathcal{N}(0, I)$; the GMM is used only to define the witness map $S(c \mid z)$.