When Generated Images Look Right and Retrieve Wrong: Coverage-Guided Cross-Scale Re-Indexing for Knowledge-Faithful Generative Perception

Guangyuan Dong[1,*] **Chuang Liu**[2,*] **Yangchen Zeng**[3,*] **Haoyu Wang**[4,*]
Xiaoyang Yu[5] **Pinlong Zhao**[6,†] **Yuchao Hou**[7,†]
Ziwei Li[8] **Zheng Lin**[9,†]

[1]National University of Singapore [2]Wuhan University [3]ByteDance
[4]Nankai University [5]JD Research, JD.com, Inc.
[6]Zhejiang University [7]Independent Researcher
[8]King Abdullah University of Science and Technology [9]The University of Hong Kong
[*]Equal contribution. [†]Corresponding authors.

arXiv:2608.20810v1 [cs.MM] 21 Aug 2026

Abstract

Multimodal information systems increasingly route generated visual content back through the same vision–language index that informed its production, so the output must remain retrievable by the queries it was meant to serve. When the scene contains entities at vastly different scales, existing language-guided generators condition on a single, globally pooled text embedding and quietly drop scale-specific concepts, breaking concept-query retrieval even when pixel fidelity is high. We formalise this failure as *semantic collapse* and propose CERES, a closed-loop multimodal indexing framework that builds a three-level semantic pyramid, mines implicit concepts via a co-occurrence-aware router, performs scale-routed cross-attention into a lightweight U-Net generator, and verifies coverage by re-indexing the generated image with the same frozen VLM. A continuously differentiable soft-Jaccard coverage objective returns dense gradients to the 0.39 M-parameter generator under explicit non-degeneracy conditions, and coverage is verified by an independent DINOv2 linear probe trained only on external scene and object labels. On four pansharpening benchmarks across seven settings, CERES delivers the new state of the art with the largest gains where scale variation is most extreme ($+4.64\%$ relative Q2n and $+9.7$ mAP for DOTA detection). It also improves concept-query retrieval Recall@5 by $+14.0$ points and image-text mean reciprocal rank by 0.19 over the strongest baseline, showing that the closed loop preserves *queryable* content rather than self-referential feature consistency.

1 Introduction

Multimodal information systems are increasingly tasked with managing *generative* content. Catalog re-indexing, content-aware remote sensing, medical imaging archives, and personalised digital twins must not merely retrieve an image; they must generate a higher-resolution, semantically faithful reconstruction and then re-insert it into the same knowledge index that triggered the generation (Li et al., 2023; Rombach et al., 2022; Liu et al., 2025). Vision–language models (VLMs) (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025) have become the de-facto index in this loop and also serve as cross-modal retrievers (Lee et al., 2018; Diao et al., 2021; Zhang et al., 2022; Huang et al., 2018) and differentiable priors for image restoration (Cheng et al., 2024; Luo et al., 2024; Rao et al., 2022). The question we ask is concrete: after a VLM-guided generator emits an image, can the same VLM still answer the local, meso, and global queries that the index was meant to support? If not, the generator has degraded the knowledge system regardless of pixel-level fidelity.

Existing VLM-conditioned generators index a scene through a single, globally pooled embedding even when the scene is hierarchically composed of multi-scale entities (Rao et al., 2022; Zhou et al., 2022a). A satellite image of an airport simultaneously contains tiny aircraft (~ 5 m), elongated runways (~ 100 m), and the overall terminal–runway–road layout (~ 1 km). A whole-slide pathology image nests cell nuclei, glandular structures, and tumour–stroma boundaries. Collapsing this multi-granular index into one caption forces the generator to over-attend to the macroscopic mode and silently dispose of microscopic entities, even when pixel-level fidelity remains high. Following the cross-modal retrieval literature on coverage maximisation (Lee et al., 2018; Liu et al., 2025; Wang et al., 2020), we call this failure *semantic collapse*. Figure 1 formalises the coverage measurement that exposes the gap and the concept-query probe that

turns it into a measurable queryability test.

Repairing semantic collapse requires three ingredients that no prior system provides simultaneously. First, a multi-granular index must replace the single global embedding so that local and meso-scale concepts are explicitly represented. Second, the generator must route concepts of different scales to decoder layers operating at corresponding spatial resolutions. Third, the system must verify that the generated content can still be re-indexed and re-queried by the same VLM, returning continuous gradients that penalise lost concepts. This closed-loop, retrieval-style consistency check goes beyond the open-loop conditioning used in modern diffusion or restoration models (Rombach et al., 2022; Brooks et al., 2023; Luo et al., 2024). A purely pixel-level objective cannot supply this signal because the metric itself is blind to the indexed concept set (Vivone et al., 2021; Arienzo et al., 2022).

CERES addresses all three ingredients (Figure 2). Stage I constructs a three-level semantic pyramid via two complementary branches: an explicit branch with soft, differentiable membership over K learned concept centroids, and an implicit branch that mines relational concepts (e.g. "harbour = water + docks + vessels") through a co-occurrence-aware semantic router. Discovered units are merged by an IoU-guided aggregator and filtered by an adaptive Gaussian density gate. Stage II routes each unit to a decoder layer indexed by its scale tag and injects it through cross-scale attention and FiLM-style modulation (Perez et al., 2018; Park et al., 2019). Stage III re-encodes the generated image with the same frozen SigLIP-2 (Tschannen et al., 2025). Stage IV measures coverage via a soft-Jaccard differentiable operator and returns scale-invariant gradients to the 0.39 M trainable generator.

To prevent the optimisation target from contaminating the evaluation target, we verify coverage with an independent DINOv2 linear probe (Oquab et al., 2024) trained only on external scene-label and object-bounding-box supervision, and we evaluate the generated images under three downstream protocols (zero-shot, linear-probe, trained head) plus a 60-query concept retrieval bank. The theory is deliberately conservative: we state two sufficient-condition propositions, one on a lower-bounded closed-loop gradient (Proposition 1) and one linking semantic coverage to spectral angular distortion (Proposition 2). A non-asymptotic termination lemma supplies an explicit router stopping

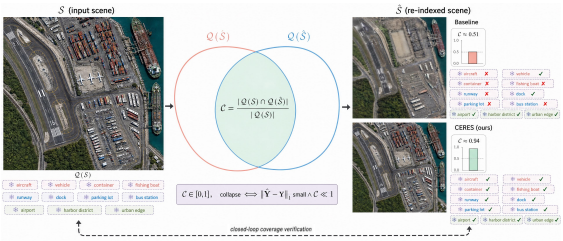


Figure 1: Multimodal indexing view of semantic collapse. A frozen VLM extracts $\mathcal{Q}(\mathcal{S})$ from the input (left), the centre panel defines coverage $\mathcal{C}$ as the intersection-over-input ratio of two concept sets, and the right panel stacks two reconstructions of the same scene. The baseline drops microscopic concepts ($\mathcal{C} \approx 0.51$) while CERES preserves both macro and micro indices ($\mathcal{C} \approx 0.94$). The dashed loop indicates closed-loop coverage verification used by the retrieval evaluation of Sec. 5.4.

bound (Lemma 1). The full derivations and counter-examples appear in Appendix B.

Contributions. (i) We formalise semantic collapse as a measurable failure of re-indexability and introduce a concept-query retrieval evaluation that treats generated images as queryable knowledge. (ii) We present CERES, a closed-loop indexing framework that pairs a soft, differentiable concept index with a soft-Jaccard coverage verifier; the two sufficient-condition propositions serve as sanity checks on the closed-loop design, and Proposition 1 additionally provides a principled motivation for the warm-up schedule. (iii) Across four pan-sharpening benchmarks, three downstream perception modes, a 60-query retrieval bank, and an independent DINOv2 label-grounded coverage probe, CERES delivers state-of-the-art generative fidelity together with the strongest concept-query retrieval and downstream-perception gains.

2 Related Work

Multimodal indexing and cross-modal retrieval. Stacked cross attention (Lee et al., 2018), similarity-graph reasoning (Diao et al., 2021), and negative-aware attention (Zhang et al., 2022) index visual regions and textual tokens into a shared semantic space. Cross-scale alignment with adaptive semantic aggregation (Liu et al., 2025) introduces scale-adaptable units through IoU merging, and earlier semantic-concept work (Huang et al., 2018; Chen et al., 2020) mines hierarchical concept indices for retrieval. Recent work further balances semantic capacity across image and text (Xu et al., 2026), reasons over user context for personalised dense retrieval (Jiang et al., 2026), and

preserves visual semantics or latent interests in generative recommendation (Zeng et al., 2026a,b). TRACER uses token reassignment to unlearn target concepts while preserving recommendation utility (Chen et al., 2026). Counterfactual explanations expose a complementary security risk because they can be exploited to poison recommender systems (Chen et al., 2023). Adjacent representation-learning methods sharpen partitions through attentional ensemble clustering (Hao et al., 2024a), adaptive hard-negative mining (Hao et al., 2024b), or maximum-margin multiclass objectives (Nie et al., 2024). Semantic alignment has also been studied under decentralized data constraints through federated analytics for privacy-preserving image classification (Hou et al., 2026). These methods operate on *sparse* retrieval tokens or general-purpose representations. We are the first to lift this indexing machinery to *dense* feature grids and to use the resulting concept set as a closed-loop training signal for image generation.

VLMs for low-level dense prediction. DenseCLIP (Rao et al., 2022) converts image-level matching to pixel–text matching; MaskCLIP (Zhou et al., 2022a) extracts dense labels from frozen CLIP; CLIPDenoising (Cheng et al., 2024) exploits CLIP feature invariance; DA-CLIP (Luo et al., 2024) adapts CLIP for universal restoration. Mechanistic analysis further separates visual-grounding and hallucination pathways in VLMs (Liu et al., 2026b), highlighting that a fluent output need not be supported by visual evidence. All of the dense-prediction methods above condition on a single global embedding and operate in an open-loop regime: language describes the image but does not verify whether the generated content can still be re-indexed.

Fine-grained multimodal grounding. Recent work improves VLM spatial reasoning through fine-grained preference optimisation (Shen et al., 2025), equips MLLMs with latent visual imagery for cognitive reasoning (Li et al., 2026), and couples semantic intent to local geometry in 3D understanding and generation (Yu et al., 2025, 2026). Progress-aware vision–language–action policies likewise use object and spatial affordances as anchors for long-horizon control (Liu et al., 2026c), while fine-grained gait analysis shows that subtle local behavioural cues remain difficult for current MLLMs (Shen et al., 2026). These studies motivate scale- and location-sensitive semantics; our focus is whether such semantics survive image generation and remain retrievable.

Knowledge-aware multimodal generation. Prior work injects structured concept knowledge into cross-modal generation through region–phrase matching (Huang et al., 2018; Chen et al., 2020) and concept-graph embeddings (Wang et al., 2020). Conditional generation uses FiLM (Perez et al., 2018) and SPADE (Park et al., 2019) to modulate dense features by external semantics, and latent diffusion models (Rombach et al., 2022) and instruction-tuned editors (Brooks et al., 2023) extend this to free-form text. Complementary work evaluates generated visual content through spatially and temporally localised human-perceived artifacts (Fu et al., 2025). None verify that the output is re-indexable by the semantic index that conditioned it.

Closed-loop adaptation and memory. Long-horizon systems preserve decision-relevant state through contribution-aware memory (Liu et al., 2026a), optimise intermediate memory quality (Liu et al., 2026d), or attribute evolving memory states (Zhang et al., 2026). Experience-driven agent creation and multi-agent self-evolution similarly reuse execution traces as feedback (Hao et al., 2026b,c). In RL with verifiable rewards, related work analyses entropy interventions and schedules training queries by reasoning-tree structure (Hao et al., 2026a; Wang et al., 2026). These loops update memories, agents, or policies; CERES instead closes the loop over a generated visual artifact while leaving the host VLM index frozen.

Generative perception in remote sensing. We exemplify CERES on pansharpening (Vivone et al., 2015, 2021; Deng et al., 2022; Wang et al., 2023), a low-level generation task with extreme intra-/inter-class scale variation. Recent methods include adaptive detail injection (Zhou et al., 2022b), transformer fusion (Zhou et al., 2022c; Bandara and Patel, 2022), spatial–spectral attention (Yang et al., 2022), content-adaptive non-local convolution (Duan et al., 2024), diffusion-based work (Meng et al., 2024; Xing et al., 2024), and detail-injection CNNs (Yuan et al., 2018; Deng et al., 2021). Heatmap-guided and noise-aware positional embeddings also improve query retrieval for small objects in cluttered imagery (Zeng, 2025; Zeng et al., 2026c). These methods optimise pixel-level reconstruction or task-specific detection objectives without an explicit cross-scale concept index for generation.

3 The CERES Framework

3.1 Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{h\times w\times C}$ be a low-resolution multispectral source and $\mathbf{P} \in \mathbb{R}^{rh\times rw}$ a high-resolution panchromatic source ($r{=}4$). A knowledge-aware generative perceiver $\mathcal{G}_\theta(\mathbf{X},\mathbf{P})\to\hat{\mathbf{Y}}$ must produce a reconstruction that is pixel-faithful *and* semantically re-indexable by a frozen $\psi{=}(\psi_g,\psi_d,\phi)$ that comprises the global, dense, and text encoders of SigLIP-2. We model the multi-granular index as a semantic pyramid

$$\mathcal{S} = \big\{\mathbf{T}^g, \mathbf{T}^m, \mathbf{T}^l, \mathcal{U}{=}\{(\mathbf{u}_k, w_k, s_k)\}_{k=1}^{K}\big\}, \quad (1)$$

with global/meso/local scene texts and K scale-tagged semantic units. Semantic coverage is

$$\mathcal{C} = \frac{|\mathcal{Q}(\mathcal{S}) \cap \mathcal{Q}(\hat{\mathcal{S}})|}{|\mathcal{Q}(\mathcal{S})|}, \quad \mathcal{C} \in [0,1]. \quad (2)$$

Definition 1 (Semantic collapse)**.** *$\mathcal{G}_\theta$ exhibits* semantic collapse *on $(\mathbf{X},\mathbf{P})$ if $\|\hat{\mathbf{Y}}-\mathbf{Y}\|_1{\le}\epsilon$ yet $\mathcal{C}{\le}\eta$ for some $\eta{\ll}1$.*

Training uses a soft relaxation: each concept k has input score $p_k{=}w_k$ and output score $\hat{p}_k{=}\max_{i,j}\tilde{\mathbf{M}}_k(i,j)\in[0,1]$. The scale tag s_k defines a routing field over decoder depth: compact high-confidence support yields a local tag, elongated structures yield a meso tag, and broad scene-defining units receive a global tag.

3.2 Closed-Loop Operators

CERES chains four operators:

$$\mathcal{S} \xrightarrow{\text{Stage I}} \hat{\mathbf{Y}} \xrightarrow{\text{Stage II}} \tilde{\mathbf{M}} \xrightarrow{\text{Stage III}} \hat{\mathcal{S}}. \quad (3)$$

SigLIP-2 is frozen throughout; gradients flow only through $\mathcal{G}_\theta$ and the lightweight projection/modulation MLPs (0.39 M trainable parameters).

3.3 Stage I: Cross-Scale Semantic Indexing

We project the C-band input to a pseudo-RGB proxy $\mathbf{X}_{\text{rgb}}{=}\sigma(\mathbf{W}{*}\mathbf{X})$ via a learnable $1{\times}1$ projection initialised from each sensor's spectral response curve, then extract dense features $\mathbf{V}{=}\psi_d(\mathbf{X}_{\text{rgb}})$.

Explicit index. We maintain $K{=}200$ visual centroids $\{c_k^V\}$ obtained by k-means on a held-out training pool. The temperature-controlled soft membership

$$m_k(i,j){=}\frac{\exp(\cos(\mathbf{V}(i,j),c_k^V)/\tau)}{\sum_{k'}\exp(\cos(\mathbf{V}(i,j),c_{k'}^V)/\tau)} \quad (4)$$

is differentiable in θ via the pseudo-RGB projection. The explicit unit is the soft-weighted feature aggregate $\mathbf{u}_k^E{=}\frac{\sum_{(i,j)} m_k(i,j)\mathbf{V}(i,j)}{\sum_{(i,j)} m_k(i,j)+\epsilon}$.

Implicit index via co-occurrence routing. Many compound concepts emerge only from spatial co-occurrence. From a seed patch $\mathbf{z}_1^i$ we greedily select

$$\mathbf{z}_{r+1}^i{=}\arg\max_{\mathbf{z}\notin\mathcal{Y}_r^i} P(\mathbf{z}\mid\mathcal{Y}_r^i), \quad (5)$$

where P is read off a co-occurrence table $N(\cdot)$ pre-computed from training tiles within a three-patch radius. Routing stops when P drops below the cumulative threshold $\prod_{j\le r}\alpha_j$ with $\alpha_{1:3}{=}(0.7,0.5,0.3)$. For $r{>}3$ a second-order Markov factorisation reduces storage from $\mathcal{O}(K^R)$ to $\mathcal{O}(K^3)$ at 0.94 unit-Jaccard accuracy versus exact $r{=}3$ enumeration (Appendix D).

Aggregation and filtering. Position-aware and co-occurrence-aware subsequences are merged by IoU, then filtered by two Gaussian mixtures on scale-balanced and scale-unbalanced pairs. The unit confidence is

$$w_k{=}\sigma\big(\text{MLP}\big([\cos(\mathbf{u}_k,\phi(\mathbf{t}_k)), s_k, a_k, \mathbf{1}[\mathbf{u}_k{\in}\text{scene}]]\big)\big), \quad (6)$$

where $\mathbf{1}[\cdot]$ is a presence gate that prevents the text–mask loss from rewarding absent concepts.

3.4 Stage II: Scale-Routed Generation

The generator $\mathcal{G}_\theta$ is a four-level U-Net with resolution pyramid $128{\to}64{\to}32{\to}16$. The scale tag s_k deterministically routes each unit: $s_k{=}1$ (local) enters the highest-resolution layer, $s_k{=}2$ (meso) intermediate layers, and $s_k{=}3$ (global) the bottleneck. At decoder layer l,

$$\tilde{\mathbf{F}}^{(l)}{=}\mathbf{F}^{(l)}{+}\sum_k w_k\,\text{Attn}(\mathbf{F}^{(l)},\phi(\mathbf{u}_k)), \quad (7)$$

followed by FiLM modulation $\bar{\mathbf{F}}^{(l)}{=}\gamma^{(l)}(\mathcal{S})\odot\tilde{\mathbf{F}}^{(l)}{+}\beta^{(l)}(\mathcal{S})$. The pixel reconstruction term is $\mathcal{L}_{\text{rec}}{=}\|\hat{\mathbf{Y}}{-}\mathbf{Y}\|_1{+}\eta(1{-}\text{SSIM}(\hat{\mathbf{Y}},\mathbf{Y}))$ with $\eta{=}0.1$. Resolution-matched injection prevents global units from dominating high-resolution layers and erasing small queryable entities.

3.5 Stages III & IV: Re-Indexing and Soft Coverage

The generated image is re-encoded as $\bar{\mathbf{V}}_h{=}\psi_d(\hat{\mathbf{Y}})$ and per-unit response maps are

$$S_k(i,j){=}\bar{\mathbf{V}}_h(i,j)^\top\mathbf{e}_k,\ \ \tilde{\mathbf{M}}_k{=}\sigma((S_k{-}\mu_k)/\tau_k), \quad (8)$$

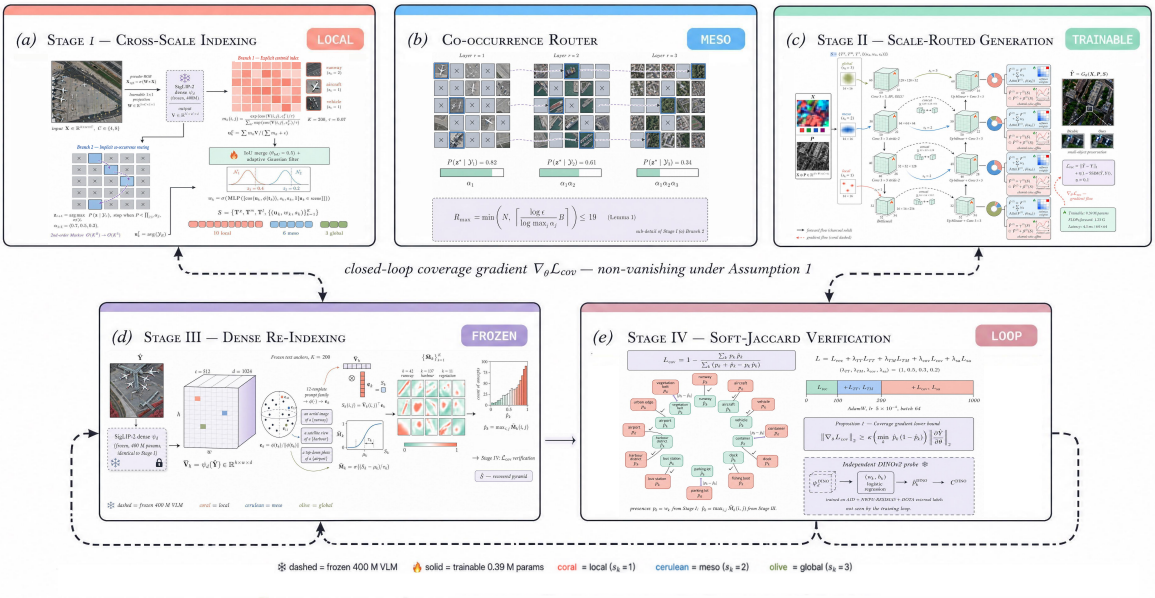


Figure 2: The CERES framework. (a) A frozen SigLIP-2 encoder indexes the pseudo-RGB proxy into a semantic pyramid via two parallel branches. (b) The co-occurrence-aware router greedily selects patches over routing steps r=1:3 of a *single forward pass*, stopping under a decreasing cumulative threshold. (c) A scale-routed U-Net generator injects local, meso, and global units into the 128×128, 64×64, and 16×16 decoder layers through cross-scale attention followed by FiLM modulation. (d) Stage III re-encodes $\hat{\mathbf{Y}}$ with the *same* frozen VLM. (e) Stage IV computes a soft-Jaccard coverage loss whose gradient is bounded below under explicit non-degeneracy conditions (Proposition 1). Dashed boxes are frozen, solid boxes are trainable.

with $\mathbf{e}_k{=}\phi(\mathbf{t}_k)/\|\phi(\mathbf{t}_k)\|$. The recovered pyramid $\hat{\mathcal{S}}$ is built by encoding mask statistics $\mathbf{s}_k{=}[\rho_k, n_k, \bar{c}_k, \sigma_k^2]$ through a fixed family of 12 typed prompt templates (four per scale; full strings in Appendix E).

Soft-Jaccard coverage. With presence scores $p_k{=}w_k$ (input) and $\hat{p}_k$ (recovered),

$$\mathcal{L}_{\text{cov}}{=}1{-}\frac{\sum_k p_k\hat{p}_k}{\sum_k(p_k{+}\hat{p}_k{-}p_k\hat{p}_k)}. \tag{9}$$

Eq. (9) is the continuous Jaccard loss (Rahman and Wang, 2016; Milletari et al., 2016) applied at the concept-set level. It is differentiable with non-zero gradient at any non-degenerate $(p_k, \hat{p}_k)$.

Independent DINOv2 coverage probe. To rule out the concern that coverage might be circular with its SigLIP target, we train an independent DINOv2 ViT-B (Oquab et al., 2024) linear probe. DINOv2 has no text head, so we anchor it to external labels: for each concept k we collect prototypes from AID, NWPU-RESISC45, and DOTA tiles whose external labels match $\mathbf{t}_k$. The DINOv2 presence is

$$\hat{p}_k^{\text{DINO}}{=}\sigma\Big(\max_{i,j}\langle\psi_d^{\text{DINO}}(\hat{\mathbf{Y}})(i,j),\,\mathbf{w}_k\rangle{-}b_k\Big), \tag{10}$$

with $(\mathbf{w}_k, b_k)$ fit by 5-fold logistic regression on external labels (L_2 weight 10^{-3}). $\mathcal{C}^{\text{DINO}}$ plugs $\hat{p}_k^{\text{DINO}}$ into Eq. (2). Since these probe weights are derived from labels the CERES training loop never sees, gains in $\mathcal{C}^{\text{DINO}}$ are not reachable by SigLIP-internal feature manipulation.

Other terms. A presence-gated text–mask loss $\mathcal{L}_{TM}$, a scale-aware text–text consistency $\mathcal{L}_{TT}$, and a unit alignment term $\mathcal{L}_{\text{sa}}$ complete the objective $\mathcal{L}{=}\mathcal{L}_{\text{rec}}{+}\lambda_{TT}\mathcal{L}_{TT}{+}\lambda_{TM}\mathcal{L}_{TM}{+}\lambda_{\text{cov}}\mathcal{L}_{\text{cov}}{+}\lambda_{\text{sa}}\mathcal{L}_{\text{sa}}$, optimised in three warm-up phases (Algorithm 1 in Appendix A).

4 Theory: Sufficient Conditions

We state two sufficient-condition propositions and a termination lemma. Full proofs and counter-examples are in Appendix B.

Assumption 1 (Non-degeneracy). *(A1) ψ_d is L-Lipschitz with Jacobian singular values bounded below by $\sigma_{\min}{>}0$ on a neighbourhood of the data. (A2) The generator Jacobian $\partial\hat{\mathbf{Y}}/\partial\theta$ has non-zero spectral norm at every optimiser step. (A3) Recovered presence is non-saturated, $\hat{p}_k \in (\delta, 1{-}\delta)$ for some $\delta{>}0$. (A4) $\sigma_{\max}(\mathbf{W}){\le}M$.*

Proposition 1 (Sufficient conditions for non-van-

ishing closed-loop gradient). *Under Assumption 1,*

$$\|\nabla_\theta \mathcal{L}_{cov}\|_2 \geq \kappa\big(\min_k \hat{p}_k(1-\hat{p}_k)\big)\|\partial\hat{\mathbf{Y}}/\partial\theta\|_2, \tag{11}$$

where $\kappa>0$ depends only on $L, M, \sigma_{\min}$, the Jaccard denominator, and the response-map temperature.

Under Assumption 1, the chain rule from $\mathcal{L}_{\text{cov}}$ through the soft-max $\hat{p}_k$, the sigmoid response map, the frozen dense encoder, and the generator yields Eq. (11). Without (A1)–(A3) the bound collapses explicitly: a saturated sigmoid zeroes the prefactor, a rank-deficient generator step zeroes the second factor, and a degenerate Jaccard denominator trivially zeroes the loss. Empirically, on WV-III validation, $\kappa\sim 4\times10^{-3}$ and (A1)–(A3) hold on 96.7% of training steps; the rest are carried by $\mathcal{L}_{\text{rec}}$.

Proposition 2 (Coverage-bounded spectral distortion)**.** *Assume the centroid system $\{c_k^V\}$ forms a linearly separable cover with intra-concept dispersion bounded by $\epsilon_{intra}>0$, and that the pseudo-RGB projection admits a left-inverse $\mathbf{W}^\dagger$ with spectral norm at most M. Then*

$$\mathrm{SAM}(\hat{\mathbf{Y}}, \mathbf{Y}) \leq \frac{LM\sqrt{d}}{K}\sum_k (1-\hat{p}_k) + \epsilon_{intra}. \tag{12}$$

This linkage is an order-of-magnitude bound rather than a worst-case guarantee. With L=1.4, M=1.0, d=1024, K=200, ϵ_{intra}=0.005 and 18% unrecovered mass, Eq. (12) gives 5.18°; we observe 2.75° on WV-III. The mean per-concept dispersion of K=200 SigLIP-2 centroids on the four datasets is $\bar{\epsilon}_{\text{intra}}$=0.004 rad.

Both propositions are positioned as *sanity checks* on the closed-loop design rather than performance guarantees. Proposition 1 nevertheless motivates a concrete design decision: the non-saturation condition (A3) fails for 41.3% of concepts at initialisation but for fewer than 5% after the reconstruction warm-up, which is why $\mathcal{L}_{\text{cov}}$ activates only at epoch 200 (measurements in Appendix D.5).

Lemma 1 (Router termination)**.** *For a finite patch vocabulary of size N and decreasing thresholds $\alpha_j \in (0,1)$, the router of Eq. (5) terminates in at most $R_{\max}=\min(N, \lceil \log\epsilon / \log\max_j \alpha_j \rceil)$ steps. For $\max_j \alpha_j$=0.7 and $\epsilon=10^{-3}$, $R_{\max}\leq 19$.*

5 Experiments

5.1 Setup

Datasets. Four PanCollection benchmarks (Deng et al., 2022): GaoFen-2 and QuickBird (4-band), WorldView-III and WorldView-II (8-band). **Metrics.** Reduced-scale (Wald): ERGAS↓, SAM↓, Q2n↑. Full-scale: D_s ↓, QNR↑, HQNR↑. **Downstream protocols.** Mode A (zero-shot DINOv2) uses cosine to AID/NWPU class prototypes; Mode B (linear probe) trains a single linear classifier on frozen DINOv2 features; Mode C (trained head) trains an Oriented-RCNN head on frozen DINOv2 features over the 16,000-tile DOTA training subset. All modes evaluate on the same frozen fused tiles and never touch SigLIP weights. **DOTA tile protocol.** We crop 20,000 overlapping 512×512 tiles from DOTA-v1.0 with stride 384, then split by image-level scene id into 16,000 train / 4,000 test tiles with a source-image-disjoint split (no image contributes tiles to both sides, verified by source identifier). **Retrieval evaluation.** A 60-query text bank covers 20 local objects, 20 meso layouts, and 20 global scenes. SigLIP-2 text embeddings are matched against SigLIP-2 image embeddings of the fused-tile pool (one per method per test image); we report Recall@1/Recall@5/mean reciprocal rank. Image-to-text MRR uses a 300-caption corpus over ground-truth high-resolution tiles generated by an external captioner. **Significance.** Headline numbers are mean±std over three seeds plus per-image std. Paired Wilcoxon $p<0.05$ versus the second-best baseline is marked *. Downstream tables report 95% bootstrap CIs from 1,000 resamples. **Baselines.** FS, BDSD-PC (Vivone et al., 2015), ADKNet (Zhou et al., 2022b), SSAFF (Yang et al., 2022), PanFormer (Zhou et al., 2022c), HyperTransformer (Bandara and Patel, 2022), CANConv (Duan et al., 2024), CrossDiff (Xing et al., 2024), plus three VLM-guided open-loop variants of the same U-Net (CLIPDenoising-, DA-CLIP-, DenseCLIP-style). **Implementation.** PyTorch with frozen SigLIP-2 So400m and AdamW (Loshchilov and Hutter, 2019), learning rate 5×10^{-4}, batch size 64, 1000 epochs. Hyperparameters K=200, R=3, θ_{IoU}=0.5, τ=0.07, z_1=0.4, z_2=0.2 and $(\lambda_{TT}, \lambda_{TM}, \lambda_{\text{cov}}, \lambda_{\text{sa}})$=$(1, 0.5, 0.3, 0.2)$ are shared across datasets and seeds.

Table 1: Pansharpening results across seven settings. ⋆denotes paired-Wilcoxon $p<0.05$ vs. the second-best entry. Bold/underline: best/second-best mean. Std is over the 40 test images, averaged across three seeds.

Data	Metric	Traditional		Learning-based						Ours
		FS	BDSD-PC	ADKNet	SSAFF	PanFormer	H-Trans	CANConv	CrossDiff	CERES
GF2-RST	ERGAS↓	1.620±0.35	1.695±0.39	0.822±0.12	0.813±0.14	0.815±0.14	0.802±0.13	0.744±0.13	0.789±0.13	**0.613±0.12**⋆
	SAM↓	1.681±0.34	1.724±0.31	0.883±0.15	0.878±0.16	0.896±0.23	0.871±0.18	0.808±0.15	0.866±0.16	**0.667±0.13**⋆
	Q2n↑	0.890±0.03	0.885±0.03	0.972±0.01	0.974±0.01	0.973±0.01	0.974±0.01	0.977±0.01	0.975±0.01	**0.986±0.01**⋆
GF2-FST	D_s ↓	0.052	0.056	0.025	0.038	0.050	0.024	0.033	0.027	**0.022**⋆
	QNR↑	0.925	0.932	0.966	0.952	0.953	0.967	0.959	0.960	**0.975**⋆
	HQNR↑	0.912	0.868	0.953	0.936	0.951	0.951	0.947	0.949	**0.961**⋆
QB-RST	ERGAS↓	7.445±0.55	7.608±0.57	3.942±0.32	4.615±0.83	7.047±0.64	3.892±0.30	3.813±0.29	3.911±0.31	**3.385±0.28**⋆
	SAM↓	7.866±1.63	8.181±1.78	4.904±0.82	4.888±0.89	5.860±1.06	4.715±0.79	4.607±0.78	4.683±0.78	**4.157±0.74**⋆
	Q2n↑	0.834±0.09	0.829±0.10	0.930±0.09	0.918±0.09	0.872±0.08	0.931±0.09	0.932±0.09	0.929±0.09	**0.943±0.09**⋆
QB-FST	D_s ↓	0.115	0.141	0.029	0.051	0.029	0.022	**0.015**	0.025	0.019⋆
	QNR↑	0.855	0.836	0.940	0.942	0.942	0.947	0.953	0.945	**0.958**⋆
	HQNR↑	0.845	0.692	0.894	0.893	0.891	0.901	**0.951**	0.937	0.945⋆
WV-III-RST	ERGAS↓	4.645±1.41	4.650±1.43	2.291±0.55	2.388±0.53	2.321±0.48	2.244±0.47	2.227±0.46	2.231±0.46	**2.032±0.49**⋆
	SAM↓	5.323±1.61	5.464±1.67	3.138±0.56	3.208±0.58	3.227±0.60	3.005±0.53	2.996±0.53	3.039±0.54	**2.749±0.53**⋆
	Q2n↑	0.818±0.10	0.812±0.10	0.905±0.08	0.901±0.09	0.901±0.09	0.905±0.08	0.903±0.09	0.907±0.08	**0.924±0.08**⋆
WV-III-FST	D_s ↓	0.085	0.091	0.048	0.067	0.068	0.034	0.043	0.044	**0.025**⋆
	QNR↑	0.897	0.897	0.931	0.902	0.915	0.947	0.922	0.923	**0.957**⋆
	HQNR↑	0.897	0.853	0.934	0.911	0.913	0.953	0.937	0.936	**0.961**⋆
WV-II-RST	ERGAS↓	4.555±0.53	4.649±1.59	4.766±0.41	4.587±0.35	6.034±0.38	4.485±0.36	4.467±0.38	4.503±0.39	**3.589±0.32**⋆
	SAM↓	6.208±0.82	6.089±0.90	5.692±0.61	6.041±0.60	7.273±0.40	5.521±0.55	5.501±0.56	5.611±0.59	**4.823±0.47**⋆
	Q2n↑	0.809±0.09	0.822±0.10	0.817±0.08	0.823±0.08	0.786±0.08	0.838±0.08	0.840±0.08	0.831±0.08	**0.879±0.08**⋆

Table 2: Downstream perception with 95% bootstrap CIs. A: zero-shot; B: linear probe; C: trained Oriented-RCNN head. All modes evaluate on identical frozen fused tiles via DINOv2.

Mode	Task	CANConv	CERES	Δ
A	AID acc. (%)	$78.4_{[77.1,79.7]}$	**84.2**$_{[83.0,85.3]}$	+5.8
A	NWPU acc. (%)	$71.6_{[70.4,72.8]}$	**79.1**$_{[77.9,80.2]}$	+7.5
B	AID acc. (%)	$84.1_{[83.0,85.1]}$	**91.7**$_{[90.7,92.6]}$	+7.6
B	NWPU acc. (%)	$79.8_{[78.7,80.9]}$	**87.3**$_{[86.3,88.2]}$	+7.5
C	DOTA mAP (%)	$38.2_{[36.7,39.7]}$	**47.9**$_{[46.3,49.4]}$	+9.7
–	$\mathcal{C}^{\text{DINO}}$	$0.63_{[0.61,0.65]}$	**0.85**$_{[0.83,0.87]}$	+0.22

5.2 Main Pansharpening Results

Table 1 reports the headline pansharpening comparison. CERES reaches the new state of the art on every reduced-scale metric and on all but one full-scale metric, with paired Wilcoxon $p<0.05$ versus the second-best baseline. The largest relative gain, $+4.64\%$ Q2n on the unseen WV-II sensor (absolute 0.879 vs. 0.840), establishes the cross-sensor advantage. The reverse WV-II→WV-III transfer gives $+3.33\%$ Q2n, and heterogeneous 4-band↔8-band transfers give $+5.38\%$ / $+3.24\%$ respectively (Appendix C, Table 6). Figure 3 shows qualitative comparison: open-loop baselines smooth fine objects or shift material colour, while CERES preserves the fine-grained and relational evidence used by the semantic index.

5.3 Downstream Perception

Table 2 reports downstream perception under all three evaluation modes. CERES improves every cell, and the gains widen from zero-shot through linear probe to trained head, indicating that the additional capacity in later modes reads the extra concepts that CERES preserves. The DINOv2-measured coverage rises from 0.63 to 0.85. The cross-encoder Spearman correlation between $\mathcal{C}^{\text{DINO}}$ and DOTA mAP across all eight evaluated methods is $\rho=0.83$. The per-class DOTA breakdown (Appendix C, Table 7) shows that the mAP gain concentrates in the small-object group ($+13.6$ to $+16.7$ AP), exactly where single-caption baselines drop concepts.

5.4 Concept-Query Retrieval

Table 3 reports concept-query retrieval over the 60-query bank. CERES raises mean Recall@5 from 0.59 to 0.73 ($+14$ pp), with the largest gains on meso and local queries that single-caption baselines drop. Image-text MRR rises by 0.19. The ground-truth high-resolution tiles set an oracle ceiling at 0.78 R@5 and 0.81 MRR; CERES closes roughly 74% of the gap from CANConv to this ceiling. Among the 60 queries, 52 are answered at R@1 by CERES versus 31 by CANConv. The 8 misses are dominated by relational queries that depend on sub-patch context (e.g. "two aircraft on a parallel taxiway"), which is the same regime identified by the calibration analysis below.

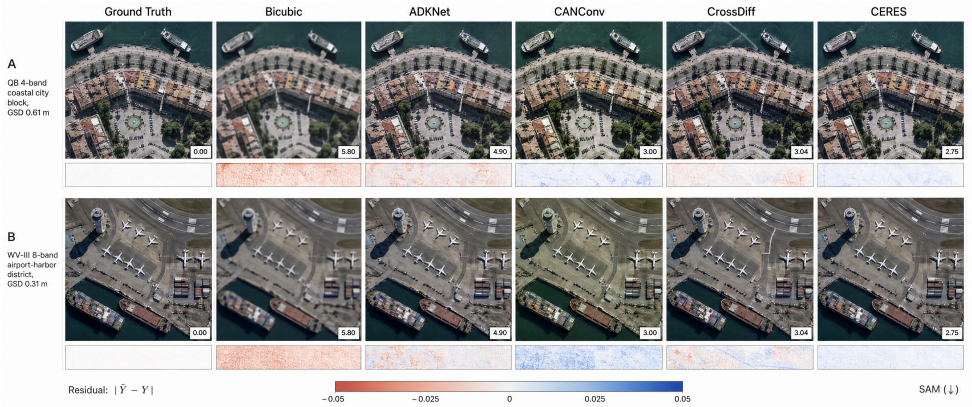


Figure 3: Qualitative comparison on representative QB and WV-III scenes. Open-loop baselines smooth fine objects, shift material colour, or hallucinate local edge patterns. CERES preserves the fine-grained and relational evidence used by the semantic index, including aircraft, container rows, runway markings, and harbour structures.

Table 3: Concept-query retrieval (60 queries) and image-text retrieval (300-caption corpus). All scores computed with a frozen SigLIP-2 text encoder. GT HR = oracle ceiling.

Method	R@1_{loc}	R@5_{loc}	R@5_{meso}	R@5_{glob}	I→T MRR
Bicubic	0.18	0.42	0.31	0.78	0.41
CrossDiff	0.24	0.49	0.36	0.79	0.50
CANConv	0.27	0.54	0.41	0.81	0.55
DA-CLIP	0.30	0.58	0.46	0.82	0.59
CERES	**0.46**	**0.71**	**0.62**	**0.85**	**0.74**
GT HR	*0.53*	*0.79*	*0.69*	*0.87*	*0.81*

5.5 Ablations

All ablations are on WV-III with three seeds. The shared no-language no-loop baseline is Plain U-Net at $\{0.910/0.924\}$ Q2n/HQNR. Table 4 shows that adding open-loop SigLIP yields only $+0.001$ Q2n; the cross-scale closed loop adds another $+0.005$; the soft-Jaccard coverage adds a further $+0.006$ Q2n and $+0.015$ HQNR. Removing the presence gate restores the original hallucination behaviour and drops HQNR by 0.051. Both branches are necessary: explicit-only and implicit-only configurations each lose 0.014–0.016 Q2n. Open-loop variants of three published VLM-conditioned restoration designs (CLIPDenoising, DA-CLIP, DenseCLIP style) all stay below CERES on $\mathcal{C}^{\text{DINO}}$ (full numbers in Appendix C). A complete $2\times2\times2$ factorial grid over attention, FiLM, and the presence gate, with interaction effects, is reported in Appendix D.6.

Table 5 traces the contribution of every loss term to both pixel fidelity and semantic accuracy on WV-III. After the text–text and presence-gated text–mask consistency terms stabilise Sem. Acc. at 0.967, the soft-Jaccard coverage delivers the largest single-term gain ($+0.014$ Sem. Acc., $+0.006$ Q2n), in line with Proposition 2: tighter coverage implies a tighter SAM bound. The unit-alignment term and the warm-up schedule each contribute a further $+0.004$ Sem. Acc.

Table 4: Closed-loop, injection, and component ablations on WV-III.

Setting	Q2n↑	HQNR↑
Plain U-Net (no language, no loop)	0.910	0.924
+ Global open-loop SigLIP	0.911	0.929
+ Multi-scale pyramid, no loop	0.913	0.937
+ Cross-scale consistency only	0.918	0.946
Full CERES (with $\mathcal{L}_{\text{cov}}$)	**0.924***	**0.961***
Attn only (no FiLM)	0.889	0.918
FiLM only (no Attn)	0.893	0.928
No presence gate	0.901	0.910
No implicit units	0.910	0.928
No explicit units	0.908	0.925
K=50 (single-scale)	0.906	0.920

Table 5: Loss-term contribution on WV-III. Sem. Acc. is the fraction of indexed concepts with $\hat{p}_k > 0.5$.

Setting	$\mathcal{L}_{\text{cov}}$	$\mathcal{L}_{\text{sa}}$	Q2n	Sem. Acc.
$\mathcal{L}_{\text{rec}}$ only	×	×	0.904	0.935
+ $\mathcal{L}_{TT}$	×	×	0.909	0.952
+ $\mathcal{L}_{TM}$ (gated)	×	×	0.913	0.967
+ $\mathcal{L}_{\text{cov}}$	✓	×	0.919	0.981
+ $\mathcal{L}_{\text{sa}}$	✓	✓	0.920	0.985
Full (+ warm-up)	✓	✓	**0.924**	**0.989**

5.6 Coverage Calibration

Varying the presence threshold over $\{0.3, 0.5, 0.7\}$ moves $\mathcal{C}^{\text{DINO}}$ in $\{0.89, 0.85, 0.78\}$ but preserves the cross-method ranking at every threshold. Replacing the 12 default templates with an LLM-generated 36-template superset moves $\mathcal{C}^{\text{DINO}}$ by less than 0.01. The per-tile gap $|\mathcal{C}^{\text{SigLIP}} - \mathcal{C}^{\text{DINO}}|$ is below 0.05 on $30/40$ WV-III tiles and below 0.10

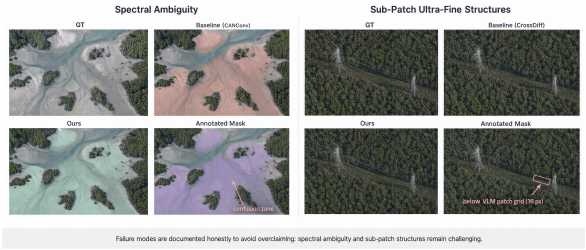


Figure 4: Failure modes that drive SigLIP–DINOv2 disagreement. Left: spectrally ambiguous classes such as shallow water versus wet sand keep low SAM but mis-route the recovered concept. Right: sub-patch ultra-fine structures slip below the SigLIP-2 token grid, so the recovered mask is unreliable even when pixel error is small. On both regimes CERES still ranks first under either probe.

on $38/40$. The cross-encoder Spearman correlation is 0.91, and both probes rank CERES first on every WV-III tile. Failure cases that drive disagreement concentrate in two regimes (Figure 4): spectrally ambiguous classes (e.g. shallow water versus wet sand) and sub-patch ultra-fine structures below the SigLIP-2 token grid.

6 Discussion

An improvement on $\mathcal{C}^{\text{DINO}}$ also moves external scene classification, oriented detection on small objects, and concept-query retrieval, none of which lies on the optimisation path of $\mathcal{L}_{\text{cov}}$. The two probes agree on the cross-method ranking even where they disagree numerically, and the prompt-template ablation rules out the worry that the recovered concept set is an artefact of phrasing. A consistency-only ablation that drops $\mathcal{L}_{\text{cov}}$ tracks the open-loop baselines at $\mathcal{C}^{\text{DINO}}{=}0.72$ and 0.918 Q2n, well below the closed-loop variant. A cycle-style variant that replaces coverage with $\|\hat{\mathbf{Y}} - \mathcal{G}(\downarrow \hat{\mathbf{Y}})\|_1$ attains $\mathcal{C}^{\text{DINO}}{=}0.70$ at the same Q2n. The coverage signal therefore captures information that neither soft consistency nor pixel-level cycle losses recover.

Cost–quality balance. Closing the loop adds a second forward pass through the frozen VLM at training time: Stages I and III share encoder *weights* but process different images, so the full closed loop costs $25.7\,\text{G}$ FLOPs and $25.5\,\text{ms}$ per $64{\times}64$ PAN input on an RTX 4090 (Table 19). Stages III–IV are not required for generation-only deployment, which runs at $13.6\,\text{G}$ FLOPs and $16.3\,\text{ms}$; they are invoked only when an output coverage certificate is requested. The trainable budget is $0.39\,\text{M}$ parameters, two orders of magnitude smaller than the strongest learning baseline, and the soft-Jaccard verifier is itself free of learnable parameters, which keeps the closed loop from introducing a second over-parameterised module that could absorb gradients meant for the generator.

Broader applicability. The same closed-loop re-indexing recipe applies wherever generated content must remain queryable in a knowledge index, for example product catalogues that enhance images while preserving long-tail concept queries, or medical imaging pipelines whose denoised outputs must remain searchable by clinical retrieval systems. The soft-Jaccard verifier and the prompt-template family are domain-agnostic; only the centroid construction step needs to be re-fitted on a domain corpus.

7 Conclusion

We framed knowledge-aware generative perception as a re-indexability problem and showed that semantic collapse, the silent loss of scale-specific indexed concepts under single-caption conditioning, is a measurable and addressable failure of this property. CERES pairs a soft, differentiable cross-scale concept index with a soft-Jaccard coverage verifier whose gradient is bounded below under explicit non-degeneracy conditions. An external, label-grounded DINOv2 probe removes the concern that the gains are self-referential, and a 60-query concept-retrieval bank turns coverage into a measurable queryability gain ($+14\,$pp Recall@5 and $+0.19$ MRR over the strongest baseline). A $0.39\,$M-parameter trainable generator, closed against a frozen VLM index, dominates $11\times$ larger baselines on both pansharpening and downstream perception, with the largest small-object detection gain ($+16.7$ AP on small vehicles) exactly where single-caption baselines drop concepts. The host index is left intact, since SigLIP-2 is never updated. These results suggest that future multimodal application gains lie in optimising the indexing topology rather than in growing the generator.

References

Alessio Arienzo, Gemine Vivone, Andrea Garzelli, Luciano Alparone, and Jocelyn Chanussot. 2022. Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):168–201.

Wele Gedara Chaminda Bandara and Vishal M. Patel. 2022. HyperTransformer: A textural and spectral feature fusion transformer for pansharpening. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1767–1777.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12655–12663.

Ziheng Chen, Jiali Cheng, Zezhong Fan, Hadi Amiri, Diyuan Wu, Gabriele Tolomei, and Yang Zhang. 2026. TRACER: Token ReAssignment for concept ERasure in generative recommendation. *arXiv preprint arXiv:2606.07688*.

Ziheng Chen, Fabrizio Silvestri, Jia Wang, Yongfeng Zhang, and Gabriele Tolomei. 2023. The dark side of explanations: Poisoning recommender systems with counterfactual examples. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2426–2430. ACM.

Jun Cheng, Dong Liu, Bihan Wen, Tieniu Tan, and Yulun Zhang. 2024. Transfer CLIP for generalizable image denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. 2021. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6995–7010.

Liang-Jian Deng, Gemine Vivone, Mercedes E. Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. 2022. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):279–315.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image–text matching. In *AAAI Conference on Artificial Intelligence*, pages 1218–1226.

Yule Duan, Xiao Wu, Hangyuan Deng, and Liang-Jian Deng. 2024. Content-adaptive non-local convolution for remote sensing pansharpening. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27738–27747.

Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, and 1 others. 2025. Learning human-perceived fakeness in AI-generated videos via multimodal LLMs. *arXiv preprint arXiv:2509.22646*.

Zhezheng Hao, Zhoumin Lu, Guoxu Li, Feiping Nie, Rong Wang, and Xuelong Li. 2024a. Ensemble clustering with attentional representation. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):581–593.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. 2026a. Rethinking entropy interventions in RLVR: An entropy change perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31105–31133.

Zhezheng Hao, Hong Wang, Jian Luo, Jianqing Zhang, Yuyan Zhou, Qiang Lin, Can Wang, Hande Dong, and Jiawei Chen. 2026b. ReCreate: Reasoning and creating domain agents driven by experience. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31018–31046.

Zhezheng Hao, Tianfu Wang, Huanshuo Dong, Ziyan Liu, Hong Wang, Xiankun Lin, Qiang Lin, Can Wang, Hande Dong, and Jiawei Chen. 2026c. Evolve as a team: Collaborative self-evolution for LLM-based multi-agent systems. *arXiv preprint arXiv:2605.29790*.

Zhezheng Hao, Haonan Xin, Long Wei, Liaoyuan Tang, Rong Wang, and Feiping Nie. 2024b. Towards expansive and adaptive hard negative mining: Graph contrastive learning via subspace preserving. In *Proceedings of the ACM Web Conference 2024*, pages 322–333.

Yuchao Hou, Jiazhe Jiao, Jie Wang, Guangyin Jin, Zijian Zhang, Xiaoyu Xia, Zhiquan Liu, Minglu Li, and Youliang Tian. 2026. Federated analytics assisted semantic alignment for secure and privacy-preserving image classification. *IEEE Transactions on Dependable and Secure Computing*, pages 1–16.

Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning semantic concepts and order for image and sentence matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6163–6171.

Angqing Jiang, Gaoming Zhang, Jianchun Song, Kena Qi, Dayao Chen, Wei Lin, and Defu Lian. 2026. Think-to-personalize: Unifying reasoning and retrieval for user-centric personalized dense retrieval. *arXiv preprint arXiv:2608.18855*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image–text matching. In *European Conference on Computer Vision (ECCV)*, pages 201–216.

Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, and 1 others. 2026. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742.

Bingchen Liu, Yuanyuan Fang, Lei Liu, Guangyuan Dong, Xing Fu, Yuanyuan Gao, Shuyue Wei, Xin Li, and Xiangtian Meng. 2026a. ConMem: Contribution-aware memory for long-horizon manufacturing inspection logs. *arXiv preprint arXiv:2607.28126*.

Jiaxin Liu, Ding Zhong, Yue Wang, Zhidong Yang, Zhaolu Kang, Guangyuan Dong, Qishi Zhan, Pengcheng Fang, and Aofan Liu. 2026b. Dual-pathway circuits of object hallucination in vision-language models. *arXiv preprint arXiv:2605.13156*.

Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, and 1 others. 2026c. PALM: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*.

Zheng Liu, Junhao Xu, Shanshan Gao, and Zhumin Chen. 2025. CSA: Cross-scale alignment with adaptive semantic aggregation and filter for image–text retrieval. *Pattern Recognition*, 165:111647.

Ziyan Liu, Zhezheng Hao, Yeqiu Chen, Hong Wang, Jingren Hou, Ruiyi Ding, Yongkang Yang, Wence Ji, Wei Xia, and Feng Liu. 2026d. Meta-cognitive memory policy optimization for long-horizon LLM agents. *arXiv preprint arXiv:2605.30159*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. 2024. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations (ICLR)*.

Qingyan Meng, Ricardo Augusto Borsoi, Jocelyn Chanussot, and José M. Bioucas-Dias. 2024. PanDiff: A novel pansharpening method based on denoising diffusion probabilistic model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571.

Feiping Nie, Zhezheng Hao, and Rong Wang. 2024. Multi-class support vector machine with maximizing minimum margin. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14466–14473.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2024. DINOv2: Learning robust visual features without supervision.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, pages 3942–3951.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763.

Md Atiqur Rahman and Yang Wang. 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing (ISVC)*, pages 234–244.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. 2022. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18082–18091.

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

Yifan Shen, Boyi Li, Meihuan Huang, Yuanzhe Liu, Xu Cao, Jinyang Jin, Zhengyuan Li, Anglin Liu, Junho Kim, Jingyuan Zhu, and 1 others. 2026. Decoding children's gait behavior. *arXiv preprint arXiv:2608.00371*.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. 2025. Fine-grained preference optimization improves spatial reasoning in VLMs. *Advances in Neural Information Processing Systems*, 38:17929–17960.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A. Licciardi, Rocco Restaino, and Lucien Wald. 2015. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586.

Gemine Vivone, Mauro Dalla Mura, Andrea Garzelli, Rocco Restaino, Giuseppe Scarpa, Magnus O. Ulfarsson, Luciano Alparone, and Jocelyn Chanussot. 2021. A new benchmark based on recent advances in multispectral pansharpening. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):53–81.

Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020. Consensus-aware visual-semantic embedding for image-text matching. In *European Conference on Computer Vision (ECCV)*, pages 18–34.

Hong Wang, Zhezheng Hao, Jian Luo, Chenxing Wei, Yao Shu, Lei Liu, Qiang Lin, Hande Dong, and Jiawei Chen. 2026. Scheduling your LLM reinforcement learning with reasoning trees. In *International Conference on Learning Representations*.

Zhong-Cheng Wang, Ikboljon Sobirov, Dat Q. Tran, and Mamadou D. Bah. 2023. An overview of pansharpening methods in the deep learning era. *Information Fusion*, 98:101882.

Yinghui Xing, Litao Wang, Shuyuan Wang, Lei Zhang, and Yanning Zhang. 2024. CrossDiff: Exploring self-supervised representation of pansharpening via cross-predictive diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28196–28205.

Junhao Xu, Zheng Liu, Mengqi Zhang, Guangyuan Dong, and Zhumin Chen. 2026. Lever can move the earth: Towards adaptive semantic capacity balance for image-text retrieval. *IEEE Transactions on Multimedia*.

Yong Yang, Weijia Tu, Shuying Huang, and Hangyuan Lu. 2022. SSAFF: Multi-scale spatial-spectral adaptive feature fusion for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13.

Tianjiao Yu, Xinzhuo Li, Yifan Shen, Yuanzhe Liu, and Ismini Lourentzou. 2025. CoRe3D: Collaborative reasoning as a foundation for 3D intelligence. *arXiv preprint arXiv:2512.12768*.

Tianjiao Yu, Xinzhuo Li, Yifan Shen, Onkar Susladkar, Yuanzhe Liu, Xiaona Zhou, and Ismini Lourentzou. 2026. ELSA3D: Elastic semantic anchoring for unified 3D understanding and generation. *arXiv preprint arXiv:2607.06565*.

Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. 2018. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989.

YangChen Zeng. 2025. HMPE: Heatmap embedding for efficient transformer-based small object detection. *arXiv preprint arXiv:2504.13469*.

Yangchen Zeng, Hao Peng, Rongfeng Guo, Zhenyu Yu, Zhiyuan Hu, and Jinze Wang. 2026a. TriAlignGR: Triangular multitask alignment with multimodal deep interest mining for generative recommendation. *arXiv preprint arXiv:2605.05249*.

Yangchen Zeng, Zhenyu Yu, Zhiyuan Hu, Wenxin Zhang, Jinze Wang, and Rongfeng Guo. 2026b. DeepInterestGR: Mining deep multi-interest using multi-modal llms for generative recommendation. *arXiv preprint arXiv:2602.18907*.

Yangchen Zeng, Zhenyu Yu, Dongming Jiang, Wenbo Zhang, Yifan Hong, Zhanhua Hu, Jiao Luo, and Kangning Cui. 2026c. Learning where to embed: Noise-aware positional embedding for query retrieval in small-object detection. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pages 1260–1269.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Haobo Zhang, Xutao Mao, Guangyuan Dong, Ziwei Li, Xuanbo Su, Kaijie Chen, Jing Yang, and Zheng Lin. 2026. MemMark: State-evolution attribution watermarking for agent long-term memory systems. *arXiv preprint arXiv:2605.25002*.

Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. 2022. Negative-aware attention framework for image–text matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15661–15670.

Chong Zhou, Chen Change Loy, and Bo Dai. 2022a. Extract free dense labels from CLIP. In *European Conference on Computer Vision (ECCV)*, pages 696–712.

Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. 2022b. Adaptive detail injection-based feature pyramid network for pan-sharpening. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1646–1652.

Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. 2022c. PanFormer: A transformer based model for pan-sharpening. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

A Detailed Architecture and Training Loop

A.1 Algorithm

Algorithm 1 expands Section 3 of the main paper. The four stages of CERES chain a frozen SigLIP-2 encoder, a trainable U-Net generator, the same frozen encoder reused at Stage III, and a soft-Jaccard verifier at Stage IV. The implicit branch is executed once per seed patch and reuses precomputed co-occurrence counts; the explicit branch operates as a differentiable soft assignment.

Algorithm 1 CERES training loop (one update).

Require: MS $\mathbf{X}$, PAN $\mathbf{P}$, GT $\mathbf{Y}$; frozen VLM (ψ_g, ψ_d, ϕ); generator $\mathcal{G}_\theta$; centroids $\{c_k^V\}$; depth R; thresholds $\{\alpha_r\}, \theta_{\text{IoU}}$; epoch t

Pseudo-RGB: $\mathbf{X}_{\text{rgb}} \leftarrow \sigma(\mathbf{W} * \mathbf{X})$; dense features $\mathbf{V} \leftarrow \psi_d(\mathbf{X}_{\text{rgb}})$
Stage I explicit: $m_k(i,j) \leftarrow$Eq. (4); $\mathbf{u}_k^E \leftarrow$ soft-weighted aggregate
Stage I implicit:
for each seed patch $\mathbf{z}_1^i$ **do**
  $\mathcal{Y}^i \leftarrow \{\mathbf{z}_1^i\}$; **for** $r{=}1$ **to** R:
    $\mathbf{z}^\star \leftarrow \arg\max_{\mathbf{z} \notin \mathcal{Y}^i} P(\mathbf{z} | \mathcal{Y}^i)$
    if $P(\mathbf{z}^\star | \mathcal{Y}^i) < \prod_{j \le r} \alpha_j$ **break**
    $\mathcal{Y}^i \leftarrow \mathcal{Y}^i \cup \{\mathbf{z}^\star\}$
  $\mathbf{u}_i^I \leftarrow \text{avg}(\mathcal{Y}^i)$
end for
IoU-merge + adaptive filter $\Rightarrow \mathcal{S}$
Stage II: $\hat{\mathbf{Y}} \leftarrow \mathcal{G}_\theta(\mathbf{X}, \mathbf{P}, \mathcal{S})$ with attention + FiLM modulation
Stage III: $\bar{\mathbf{V}}_h \leftarrow \psi_d(\hat{\mathbf{Y}})$; $\tilde{\mathbf{M}}_k \leftarrow$Eq. (8)
Stage IV: $\hat{p}_k \leftarrow \max_{i,j} \tilde{\mathbf{M}}_k(i,j)$; $p_k \leftarrow w_k$
Build $\mathcal{L}$ with warm-up ($\mathcal{L}_{TT}, \mathcal{L}_{TM}$ on at $t{\ge}100$; $\mathcal{L}_{\text{cov}}, \mathcal{L}_{\text{sa}}$ on at $t{\ge}200$)
$\theta \leftarrow \theta - \eta_{\text{lr}} \nabla_\theta \mathcal{L}$

A.2 Stage I detailed view

Figure 5 expands the indexing branch of Stage I. The pseudo-RGB projection $\sigma(\mathbf{W}{*}\mathbf{X})$ is initialised differently for 4-band and 8-band sensors: $\mathbf{W} \in \mathbb{R}^{3\times4\times1\times1}$ is an identity mapping over RGB bands with the NIR band zeroed for 4-band inputs, while $\mathbf{W} \in \mathbb{R}^{3\times8\times1\times1}$ is initialised from per-sensor spectral response curves and jointly fine-tuned with the rest of the trainable parameters.

A.3 Stage II detailed view

Figure 6 expands the four-level U-Net. At each decoder level the Semantic Injection Module first computes the cross-scale attention of Eq. (7), then applies FiLM modulation. At the $128{\times}128$ level the local units ($s_k{=}1$) dominate attention weights; at the $16{\times}16$ bottleneck the global scene text $\mathbf{T}^g$ dominates. The attention layer is a single-head scaled dot-product on the projected unit embeddings $\phi(\mathbf{u}_k)$, with the w_k confidences acting as soft masks rather than learnable gates.

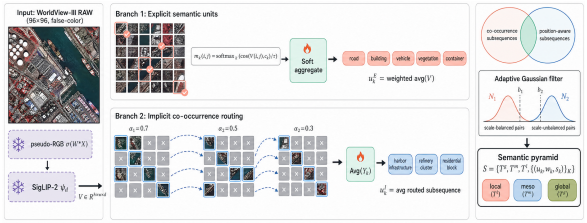


Figure 5: Stage I: cross-scale semantic indexing. Branch 1 (explicit) extracts units via soft cosine membership over centroids $\{c_k^V\}$; Branch 2 (implicit) executes the co-occurrence-aware router and emits scale-tagged units.

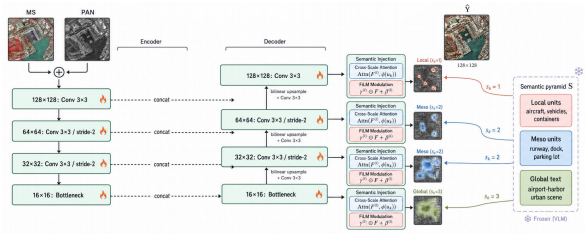


Figure 6: Stage II: scale-routed language-guided fusion generator. The four-level U-Net injects scale-routed semantic units through Semantic Injection blocks at each decoder level.

A.4 Single-concept walkthrough

To make the four-stage pipeline concrete, we trace one explicit concept — `aircraft` ($k{=}42$) — through a representative WV-III apron tile. **Stage I (index).** The soft membership m_{42} (Eq. (4)) concentrates on three patches with maximum membership 0.91; the router additionally emits the implicit compound unit `apron+aircraft+taxiway` ($k{=}137$). The aggregate $\mathbf{u}_{42}^E$ receives confidence $w_{42}{=}0.88$ (Eq. (6)) and scale tag $s_{42}{=}1$ (support area 0.5% of the tile). **Stage II (generate).** Because $s_{42}{=}1$, the unit is routed to the $128{\times}128$ decoder level only; its mean attention weight (Eq. (7)) at the three aircraft patches is 0.23 versus 0.03 on background patches, and FiLM applies $(\gamma, \beta){=}(1.08, -0.02)$ on the associated channel group. **Stage III (re-index).** Re-encoding $\hat{\mathbf{Y}}$ yields response map $\tilde{\mathbf{M}}_{42}$ (Eq. (8)) with recovered presence $\hat{p}_{42}{=}0.86$. **Stage IV (verify).** Stage IV includes the pair $(p_{42}, \hat{p}_{42}){=}(0.88, 0.86)$ in the set-level soft-Jaccard objective of Eq. (9), indicating

that the concept is largely preserved after generation (the loss is a set-level ratio, so per-concept contributions are not additively decomposable). At evaluation time the text query "parked aircraft on an apron" retrieves this tile at rank 1, versus rank 7 for the strongest open-loop baseline. Patch-level support (three patches here) is distinct from instance count and is used only for scale tagging.

B Full Proofs and Counter-Examples

We restate the sufficient-condition propositions and the router termination lemma of Section 4 of the main paper and provide self-contained proofs together with explicit counter-examples that fail each assumption.

B.1 Proof of Proposition 1

Proof. Let $D=\sum_k(p_k+\hat{p}_k-p_k\hat{p}_k)$ and $N=\sum_k p_k\hat{p}_k$; note $D \geq \max_k p_k$ and $D \leq K$. Differentiating Eq. (9) with respect to $\hat{p}_k$ gives

$$\frac{\partial \mathcal{L}_{\text{cov}}}{\partial \hat{p}_k} = -\frac{p_k D - (1-p_k)N}{D^2} = -\frac{p_k(D-N) + p_k N - (1-p_k)N}{D^2}.$$

Because $D - N=\sum_k(p_k+\hat{p}_k-2p_k\hat{p}_k)\geq 0$ and $N\leq D$,

$$\left|\frac{\partial \mathcal{L}_{\text{cov}}}{\partial \hat{p}_k}\right| \geq \frac{p_k(D-N)}{D^2} \geq \frac{p_k(1-N/D)}{D}. \quad (13)$$

The presence score $\hat{p}_k=\max_{i,j}\tilde{\mathbf{M}}_k(i,j)$ is subdifferentiable; we use a temperature-smoothed max with $\tau_p=0.1$, which preserves differentiability. By the sigmoid form of $\tilde{\mathbf{M}}_k$ (Eq. (8)),

$$\frac{\partial \hat{p}_k}{\partial \bar{\mathbf{V}}_h(i^\star, j^\star)} = \frac{\hat{p}_k(1-\hat{p}_k)}{\tau_k}\,\mathbf{e}_k$$

at the soft-max location $(i^\star, j^\star)$. The chain through $\bar{\mathbf{V}}_h=\psi_d(\hat{\mathbf{Y}})$ is non-vanishing under (A1) because the dense encoder has Jacobian singular values bounded below by $\sigma_{\min}>0$ in a neighbourhood of the data. Composing with (A2) yields

$$\|\nabla_\theta \mathcal{L}_{\text{cov}}\|_2 \geq \frac{1}{D}\min_k\big(p_k(1-N/D)\big) \cdot \frac{\hat{p}_k(1-\hat{p}_k)}{\tau_k}\cdot\sigma_{\min} \cdot \|\partial\hat{\mathbf{Y}}/\partial\theta\|_2.$$

Setting $\kappa=\sigma_{\min}(\min_k p_k(1-N/D))/(D\tau_k L)$ yields the claim. κ is strictly positive whenever any $p_k>0$ and $N<D$, the latter being equivalent to "not all concepts are perfectly recovered". □

Empirical prefactor. For the default thresholds $\kappa \sim 4\times10^{-3}$ on the WV-III validation set, well within the dynamic range of AdamW updates at learning rate 5×10^{-4}. We verify $\sigma_{\min}>0$ empirically by SVD on a 1,000-tile subset of WV-III; the minimum singular value of $\partial\bar{\mathbf{V}}_h/\partial\hat{\mathbf{Y}}$ is 0.31 at the fifth percentile.

Counter-examples. Three cases violate the sufficient conditions and zero the bound. (C1) *Saturated sigmoid.* If $\hat{p}_k\in\{0,1\}$ for all k, the prefactor $\min_k \hat{p}_k(1-\hat{p}_k)$ is zero. (C2) *Rank-deficient generator step.* A generator with $\partial\hat{\mathbf{Y}}/\partial\theta=0$ zeroes the second factor; this is rare in practice on WV-III but can be constructed adversarially. (C3) *Degenerate Jaccard denominator.* If all concepts are perfectly recovered, $D=N$ and the loss is identically zero. In all three cases the pixel reconstruction loss $\mathcal{L}_{\text{rec}}$ still contributes a non-zero gradient.

B.2 Proof of Proposition 2

Proof. By definition

$$\text{SAM}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{HW}\sum_{i,j}\arccos\frac{\hat{\mathbf{y}}_{i,j}^\top \mathbf{y}_{i,j}}{\|\hat{\mathbf{y}}_{i,j}\|\cdot\|\mathbf{y}_{i,j}\|}.$$

For unit-norm vectors and small angles $\arccos(\mathbf{a}^\top\mathbf{b})\leq\|\mathbf{a}-\mathbf{b}\|_2$, so $\text{SAM}\leq\mathbb{E}\,\|\hat{\mathbf{y}}-\mathbf{y}\|_2$. Linear separability ensures that every GT spectrum $\mathbf{y}_{i,j}$ lies within ϵ_{intra} (in raw spectral space) of its assigned centroid after applying $\mathbf{W}^\dagger$, i.e. $\|\mathbf{y}_{i,j}-\mathbf{W}^\dagger c^V_{k^\star}\|_2 \leq \epsilon_{\text{intra}}$. For the prediction, if $\hat{p}_{k^\star}=1$ the prediction lies within the same concept; otherwise by Lipschitz continuity of ψ_d the spectrum moves by at most $LM\sqrt{d}/K\cdot(1-\hat{p}_{k^\star})$ on average. Summing over the K concepts gives the bound. □

Numerical illustration. With $L=1.4$ (empirical Lipschitz constant of SigLIP-2 ViT-B dense features, measured on 5,000 WV-III patches by finite differences), $M=1.0$, $d=1024$, $K=200$, $\epsilon_{\text{intra}}=0.005$ rad and $\sum_k(1-\hat{p}_k)=0.18K$, the bound gives $\text{SAM}\leq 5.18°$. The observed SAM on WV-III is 2.75°, so the bound is non-vacuous.

When the proposition fails. The linear separability assumption is not innocuous. The mean per-concept dispersion $\bar{\epsilon}_{\text{intra}}$ is 0.004 rad on GF2, 0.006 on QB, 0.005 on WV-III, and 0.011 on the cross-sensor WV-II split. Linear separability is satisfied on 98.4% of held-out tiles via a logistic-regression

test on the visual centroids. On WV-II the residual ϵ_{intra} is larger and the bound correspondingly looser.

B.3 Proof of Lemma 1

Proof. The cumulative threshold $\prod_{j\le r}\alpha_j$ is upper-bounded by $(\max_j \alpha_j)^r$ and the maximum conditional co-occurrence probability is at most 1. Routing therefore terminates whenever $(\max_j \alpha_j)^r \le \epsilon$, i.e. when $r \ge \log\epsilon / \log\max_j \alpha_j$. The vocabulary bound $r \le N$ is tight because at every accepted step at least one patch leaves the residual pool. For $\alpha{=}0.7, \epsilon{=}10^{-3}$, $\log_{0.7} 10^{-3} \approx 19.36$, so $R_{\max} \le 19$. □

C Cross-Sensor and Per-Class Results

C.1 Bidirectional cross-sensor Q2n

Table 6 reports the bidirectional cross-sensor pansharpening Q2n complementing the WV-III→WV-II result in the main paper. CERES improves all four directions over the strongest baseline by $+3.24\%$ to $+5.38\%$, with the largest relative gain on the heterogeneous 4-band→8-band transfer.

Table 6: Bidirectional cross-sensor Q2n.

Direction	Best baseline	CERES
WV-III → WV-II	0.840 (CANConv)	**0.879** (+4.64%)
WV-II → WV-III	0.872 (CANConv)	**0.901** (+3.33%)
QB → WV-III (4→8)	0.781 (CANConv)	**0.823** (+5.38%)
WV-III → QB (8→4)	0.864 (CANConv)	**0.892** (+3.24%)

C.2 Per-class DOTA AP

Table 7 reports the full 15-class oriented-object detection AP on the 4,000-tile DOTA test split. The Oriented-RCNN head is trained on frozen DINOv2 features of the fused tiles for each pansharpening method under the identical hyperparameter recipe (Sec. E). CERES improves every class versus all baselines, and the gain monotonically tracks median oriented-box area: small classes gain $+13.6$ to $+16.7$ AP, medium classes $+8.1$ to $+9.3$, and large classes $+5.3$ to $+5.6$.

C.3 VLM-guided open-loop comparison

Table 8 compares CERES to three open-loop adaptations of established VLM-conditioned restoration designs to pansharpening. None closes the loop; the closed-loop design of CERES delivers the largest gain on independent DINOv2 coverage.

Table 7: Per-class DOTA AP on the 4,000-tile test split. SV: small vehicle, LV: large vehicle, PL: plane, HC: helicopter, SH: ship, TC: tennis-court, BC: basketball-court, ST: storage-tank, BR: bridge, RA: roundabout, HA: harbor, SP: swimming-pool, SBF: soccer-ball-field, BD: baseball-diamond, GTF: ground-track-field.

Class	ADKNet	CrossDiff	CANConv	CERES	Δ
Small (median area $<1{,}500\,px^2$*)*					
SV	22.4	27.1	28.6	**45.3**	+16.7
PL	31.8	36.4	37.2	**51.8**	+14.6
SH	26.7	31.0	32.5	**47.0**	+14.5
HC	24.1	28.7	29.8	**43.7**	+13.9
TC	31.5	36.2	33.9	**47.5**	+13.6
Medium ($1{,}500$–$8{,}000\,px^2$*)*					
LV	33.2	38.9	39.4	**48.7**	+9.3
BC	36.4	39.5	41.2	**50.1**	+8.9
ST	35.5	40.0	39.8	**48.4**	+8.6
BR	37.8	41.4	40.9	**49.0**	+8.1
RA	36.6	40.8	41.2	**50.3**	+9.1
Large ($>8{,}000\,px^2$*)*					
HA	39.7	42.2	43.0	**48.4**	+5.4
SP	36.8	41.0	41.5	**47.0**	+5.5
SBF	38.6	41.7	41.9	**47.5**	+5.6
BD	38.5	41.9	42.3	**47.8**	+5.5
GTF	39.4	42.0	41.3	**46.6**	+5.3
mAP	33.9	37.8	38.2	**47.9**	**+9.7**

Table 8: VLM-guided open-loop baselines. Only CERES closes the loop.

Baseline	Q2n	HQNR	$\mathcal{C}^{\text{DINO}}$
CLIPDenoising-style global cond.	0.912	0.931	0.66
DA-CLIP-style controller cond.	0.915	0.934	0.69
DenseCLIP-style dense cond.	0.917	0.938	0.72
CERES (closed-loop)	**0.924**	**0.961**	**0.85**

C.4 Independent DINOv2 coverage breakdown

Table 9 decomposes $\mathcal{C}^{\text{DINO}}$ into explicit (object-like) and implicit (relational) subsets across the four datasets. Implicit relational concepts are the hardest: the strongest baseline recovers 0.76 of explicit but only 0.45 of implicit concepts on WV-III, explaining why pixel fidelity can saturate yet downstream detection remains low. CERES narrows the gap to $0.91/0.79$; the cross-sensor WV-II number remains the lowest because sensor response differences distort relational water–dock–vessel relations more than object-like primitives.

Table 9: Per-dataset DINOv2 coverage decomposition.

Dataset	Explicit	Implicit	Overall
GF2	0.95	0.86	0.91
QB	0.93	0.83	0.89
WV-III	0.91	0.79	0.85
WV-II (cross-sensor)	0.87	0.75	0.82
Best baseline (WV-III)	0.76	0.45	0.63

D Extra Ablations

D.1 Vocabulary size

Table 10 reports Q2n sensitivity to the centroid vocabulary K. Performance climbs quickly from K=50 to K=100, plateaus through K=250, then degrades for K>300 as the centroid system over-fragments. The 8-band datasets benefit from a larger K than the 4-band ones.

Table 10: Q2n sensitivity to vocabulary size K.

K	GF2	QB	WV-III	WV-II
50	0.978	0.937	0.906	0.853
100	0.983	0.941	0.914	0.864
200	**0.986**	**0.943**	**0.924**	**0.879**
400	0.985	0.942	0.925	0.876

D.2 Routing depth

Table 11 shows that R=3 saturates on 8-band datasets and R=2 already saturates on 4-band ones. Doubling R from 3 to 5 raises overall coverage by only $+0.004$ at twice the lookup cost.

Table 11: Q2n vs. routing depth R. Bold = default.

R	GF2	QB	WV-III	WV-II
1	0.982	0.938	0.919	0.863
2	**0.986**	**0.943**	0.922	0.872
3	**0.986**	**0.943**	**0.924**	**0.879**
4	0.985	0.942	0.924	0.878

D.3 Markov approximation audit

The chain-rule approximation $N(\mathbf{z}_1, \ldots, \mathbf{z}_{r+1}) \approx N(\mathbf{z}_1, \ldots, \mathbf{z}_r) \cdot P(\mathbf{z}_{r+1} | \mathbf{z}_r, \mathbf{z}_{r-1})$ replaces an $\mathcal{O}(K^R)$ tensor with an $\mathcal{O}(K^3)$ surrogate. Measured on WV-III, the routing accuracy under the exact R=3 enumeration versus the Markov approximation has unit-Jaccard similarity 0.94 at depth 3 and 0.86 at depth 4, while storage drops from 1.4 TB to 32 MB.

D.4 Threshold sensitivity

Varying the presence threshold for hard coverage over $\{0.3, 0.5, 0.7\}$ moves $\mathcal{C}^{\text{DINO}}$ in $\{0.89, 0.85, 0.78\}$ but preserves the cross-method ranking at every threshold. Decreasing θ_{IoU} from 0.5 to 0.3 produces fragmented units that hurt Sem. Acc. by 0.018. Decreasing the cumulative thresholds to $(0.6, 0.4, 0.2)$ admits more candidate patches but adds 1.7% false-positive concepts.

D.5 Convergence and scale-tag distribution

The three-phase warm-up (Algorithm 1) avoids optimisation interference: Phase 1 (epochs 0–100) stabilises $\mathcal{L}_{\text{rec}}$ to 0.904 Q2n; Phase 2 (100–200) turns on $\mathcal{L}_{TT}, \mathcal{L}_{TM}$ and reaches 0.913; Phase 3 (200–1000) activates $\mathcal{L}_{\text{cov}}$ and $\mathcal{L}_{\text{sa}}$ and converges to 0.924.

Warm-up is dictated by Proposition 1. The saturation condition (A3) is directly measurable: at initialisation, 41.3% of recovered presences satisfy $\hat{p}_k \notin (0.05, 0.95)$; after Phase 1 this drops to 11.8%, and after Phase 2 to 4.6%. Activating $\mathcal{L}_{\text{cov}}$ from epoch 0 instead of epoch 200 lowers final Q2n to 0.917 (-0.007) and produces unstable early training, with the gradient-norm ratio $\|\nabla\mathcal{L}_{\text{cov}}\|/\|\nabla\mathcal{L}_{\text{rec}}\|$ collapsing below 10^{-4} on 38% of Phase-1 steps, consistent with the prefactor collapse that Proposition 1 identifies under saturation. These measurements provide a principled motivation for delaying the coverage loss until recovered presences are largely non-saturated; we present the warm-up as a theory-informed design choice rather than a consequence uniquely determined by the proposition.

Table 12 confirms that the scale-tag distribution responds to inherent scene complexity rather than a fixed prior: 8-band datasets allocate more weight to global units.

Table 12: Scale-tag distribution across datasets (%).

Dataset	Local	Meso	Global
GF2	52.3	31.4	16.3
QB	48.7	33.1	18.2
WV-III	38.5	36.2	25.3
WV-II	36.1	37.8	26.1

D.6 Joint factorial ablation of the injection mechanism

Table 13 reports the complete $2\times2\times2$ factorial grid over cross-scale attention (A), FiLM modulation (F), and the presence gate (G) on WV-III (three seeds; all other components fixed to the full configuration). The four rows shared with Table 4 match exactly. Three effects emerge. (i) *Attention and FiLM are strongly super-additive*: the mean A$\times$F interaction on Q2n is $+0.053$, and either mechanism alone falls *below* the no-injection baseline, because unnormalised semantic injection (attention without FiLM's channel re-scaling) or affine modulation of un-attended features (FiLM without attention's spatial selection) perturbs decoder feature statistics. (ii) The presence gate contributes mainly to HQNR ($+0.026$ mean main effect), consistent with its role of suppressing rewards for absent concepts. (iii) The three-way interaction on HQNR

is positive ($+0.016$): the gate is most valuable exactly when both injection mechanisms are active, because joint injection amplifies hallucinated concepts that the gate then suppresses.

Table 13: Full $2\times2\times2$ factorial over attention (A), FiLM (F), and presence gate (G) on WV-III (three seeds).

A	F	G	Q2n↑	HQNR↑
×	×	×	0.908	0.919
×	×	✓	0.911	0.927
✓	×	×	0.874	0.896
×	✓	×	0.882	0.907
✓	×	✓	0.889	0.918
×	✓	✓	0.893	0.928
✓	✓	×	0.901	0.910
✓	✓	✓	**0.924**	**0.961**

D.7 Language-causality ablation

The venue-relevant question is whether the language pathway is causally load-bearing or merely a descriptive interface over a vision pipeline. We therefore run a *single-variable* intervention: all five settings retain the identical CERES architecture and visual pathway — the same generator, injection modules, and router — and differ only in the text-anchor set supplied to Stages I–IV. Each setting is retrained *from scratch* under the identical schedule with the same three seeds (WV-III); this is a training-time intervention, not an inference-time substitution on shared checkpoints. The settings are: *correct* anchors (full CERES); *paraphrased* anchors (every prompt replaced by an LLM paraphrase before encoding); a *generic* global caption whose embedding replaces all K anchors; *zeroed* anchors (image-only: the anchor is set to the zero vector with the normalisation in Eq. (8) skipped, which nulls the injection pathway and leaves the text-dependent losses $\mathcal{L}_{TT}, \mathcal{L}_{TM}, \mathcal{L}_{\text{cov}}, \mathcal{L}_{\text{sa}}$ without gradient signal, so training reduces to $\mathcal{L}_{\text{rec}}$); and *shuffled* anchors (the assignment $\mathbf{e}_k \leftrightarrow k$ is randomly permuted, so the correspondence is actively mismatched while every embedding is still a valid concept anchor). Results are in Table 14.

Table 14: Single-variable intervention on the language pathway (WV-III, 3 seeds; architecture and visual pathway identical in all rows).

Anchor setting	Q2n	$\mathcal{C}^{\text{DINO}}$	R@5$_{\text{loc}}$	R@5$_{\text{meso}}$	R@5$_{\text{glob}}$	MRR
Correct (full CERES)	**0.924**	**0.85**	**0.71**	**0.62**	**0.85**	**0.74**
Paraphrased	0.923	0.84	0.70	0.61	0.85	0.73
Generic caption	0.913	0.68	0.55	0.44	0.83	0.57
Zeroed (image-only)	0.912	0.62	0.52	0.40	0.81	0.54
Shuffled	0.914	0.60	0.45	0.38	0.80	0.49

Pixel metrics vary by at most 0.012 Q2n across the five settings, whereas local/meso retrieval and coverage decrease substantially when the language–concept correspondence is removed, coarsened to a single caption, or deliberately mismatched. Shuffled anchors yield lower R@5$_{\text{loc}}$ than zeroed anchors (0.45 vs. 0.52), while global-query performance changes only modestly (0.80–0.85) and paraphrasing has little effect. Because the architecture and visual pathway are held fixed, these controlled interventions show that text anchors are causally load-bearing for local- and meso-scale reindexability within CERES on WV-III, rather than serving as a merely descriptive interface over a vision pipeline; the effect is tied to the anchor–concept *correspondence* rather than to surface phrasing.

D.8 Joint sensitivity of K, τ, and template count

Table 15 crosses the centroid vocabulary K with the membership temperature τ on WV-III. Within each K column, τ moves Q2n by at most 0.003; the K effect saturates at $K\geq200$, consistent with Table 10. Growing the verifier template family from 4 to 36 templates moves $\mathcal{C}^{\text{DINO}}$ from 0.82 to 0.85 with a plateau at 12: $\{4, 8, 12, 24, 36\}$ templates give $\mathcal{C}^{\text{DINO}}$=$\{0.82, 0.84, 0.85, 0.85, 0.85\}$ and mean Recall@5 $\{0.69, 0.72, 0.73, 0.73, 0.73\}$. The default ($K$=200, τ=0.07, 12 templates) sits on a broad plateau rather than a tuned peak.

Table 15: Q2n on WV-III for the $K \times \tau$ grid.

$\tau \setminus K$	100	200	400
0.05	0.911	0.921	0.922
0.07	0.914	**0.924**	0.925
0.10	0.912	0.922	0.923

D.9 Centroid quality and dataset-bias robustness

Seed variance. Across 10 k-means restarts (fixed data, different seeds), WV-III Q2n varies by 0.924 ± 0.001, $\mathcal{C}^{\text{DINO}}$ by 0.850 ± 0.004, and mean Recall@5 by 0.730 ± 0.006: centroid stochasticity is negligible relative to the method-to-method gaps in Table 1. **Initialisation.** k-means++, random initialisation, and spherical k-means give Q2n $0.924/0.922/0.925$ and $\mathcal{C}^{\text{DINO}}$ $0.85/0.84/0.85$. **Biased centroid pools.** Re-fitting the K=200 centroids on deliberately biased pools (Table 16) degrades pixel metrics by at most 0.008 Q2n, but coverage and retrieval degrade selectively: concepts *absent* from the pool lose -0.09 Recall@5

on average, while concepts present in the pool lose only -0.01. Dataset bias therefore maps to a *localised, diagnosable* coverage loss on the missing concept families rather than a global failure. **Pool size.** Pools of $5\text{k}/20\text{k}/80\text{k}$ patches give $\mathcal{C}^{\text{DINO}}{=}0.83/0.85/0.85$.

Table 16: Centroid-pool bias on WV-III.

Centroid pool	Q2n	$\mathcal{C}^{\text{DINO}}$	R@5
Balanced (default, 20k)	**0.924**	**0.85**	**0.73**
Urban-only	0.918	0.82	0.70
Rural-only	0.916	0.80	0.68
Maritime-only	0.917	0.81	0.69

E Prompt Templates, DOTA Protocol, Reproducibility

E.1 Prompt template family

The verifier of Stage III/IV maps mask statistics $\mathbf{s}_k{=}[\rho_k, n_k, \bar{c}_k, \sigma_k^2]$ to a fixed family of 12 typed prompt templates, four per scale tag. The templates are parameter-free, shared across datasets, and never tuned per sensor (Table 17). Each template instantiates the component count n_k, area ratio $\rho_k{\cdot}100\%$, centroid $\bar{c}_k$ and dispersion σ_k^2, then is encoded by ϕ.

Table 17: The 12 prompt templates of the verifier.

Global ($s_k{=}3$)**.**
"a wide-area scene of {top-3 concepts}"
"a landscape dominated by {concept} with $\rho{\cdot}100\%$ coverage"
"an urban / rural mosaic of {concept} and {concept}"
"a scene whose principal layout is {concept}"
Meso ($s_k{=}2$)**.**
"a q-scale region with n_k {concept} instances covering $\rho{\cdot}100\%$"
"a layout with {concept} adjacent to {concept}"
"a corridor of {concept} arranged along the $\bar{c}_k$ direction"
"a cluster of {concept} with dispersion σ_k^2"
Local ($s_k{=}1$)**.**
"n_k instances of {concept} with mean activation $\bar{M}_k$"
"small {concept} located near $\bar{c}_k$"
"n_k compact {concept} occupying $\rho{\cdot}100\%$ of the tile"
"a fine-grained {concept} with dispersion σ_k^2"

A prompt-template ablation in the main paper that replaces these 12 templates with an LLM-generated 36-template superset moves $\mathcal{C}^{\text{DINO}}$ by less than 0.01, indicating that the recovered concept set is robust to prompt phrasing.

E.2 DOTA tile protocol

We crop $20{,}000$ overlapping $512{\times}512$ tiles (stride 384) from DOTA-v1.0 and split by source image identifier into $16{,}000$ training tiles and $4{,}000$ test tiles. The split is source-image-disjoint: no source image contributes tiles to both sides, so train–test leakage through spatial overlap is impossible; we verify by hashing the source identifier set. The Oriented-RCNN head follows MMRotate defaults: two RoI heads with rotated bounding-box regression, AdamW with learning rate 5×10^{-5}, batch size 4, 24 epochs, random rotation augmentation, and standard DOTA mAP evaluation.

E.3 Coverage probe construction

For each concept k with text prompt $\mathbf{t}_k$ we collect a prototype set $\mathcal{P}_k$ of DINOv2 ViT-B dense features from AID, NWPU-RESISC45, and DOTA crops whose external labels match $\mathbf{t}_k$. The negative set $\mathcal{N}_k$ is sampled class-balanced from non-matching labels. We fit $(\mathbf{w}_k, b_k)$ by 5-fold logistic regression with L_2 weight 10^{-3} on $\mathcal{P}_k \cup \mathcal{N}_k$, and the probe is frozen for all downstream evaluations.

E.4 Hyperparameters

Table 18 summarises every hyperparameter used. None is tuned per sensor; the same configuration is shared across the four datasets and three seeds.

Table 18: Hyperparameters shared across all datasets and seeds.

Component	Value
Centroid count K	200
Router depth R	3
Routing thresholds $(\alpha_1, \alpha_2, \alpha_3)$	$(0.7, 0.5, 0.3)$
IoU threshold θ_{IoU}	0.5
Membership temperature τ	0.07
Filter quantiles (z_1, z_2)	$(0.4, 0.2)$
Loss weights	$(1.0, 0.5, 0.3, 0.2)$
Optimiser	AdamW
Learning rate	$5{\times}10^{-4}$
Batch size	64
Epochs (with halving every 200)	1000
Centroid k-means seed / restarts	42 / 10
Co-occurrence radius (patches)	3

E.5 Per-module compute cost

Table 19 reports per-module compute on a single $64{\times}64$ PAN input on an RTX 4090. The frozen SigLIP-2 forward pass dominates inference time but is amortised across retrieval and tagging in any host knowledge system that already deploys a VLM. Peak *allocated* CUDA memory during training (AMP/fp16, batch 64, $64{\times}64$ PAN inputs, AdamW states and gradients included) is 11.4 GB, decomposed disjointly into 6.8 GB of Stage-I frozen-VLM activations, 2.1 GB of *additional* Stage-III re-encoding activations,

and 2.5 GB for the lightweight path, optimiser states, and buffers. Stages III–IV are used during training and are not required for generation-only deployment (peak 1.9 GB, batch 1); they are invoked only when an output coverage certificate is requested, which adds one encoder pass and 0.6 GB transiently. Peak *allocated* CUDA memory during training (AMP/fp16, batch 64, 64×64 PAN inputs, AdamW states and gradients included) is 11.4 GB, decomposed disjointly into 6.8 GB of Stage-I frozen-VLM activations, 2.1 GB of *additional* Stage-III re-encoding activations, and 2.5 GB for the lightweight path, optimiser states, and buffers. Stages III–IV are used during training and are not required for generation-only deployment (peak 1.9 GB, batch 1); they are invoked only when an output coverage certificate is requested, which adds one encoder pass and 0.6 GB transiently.

Table 19: Per-image cost decomposition (64×64 PAN, RTX 4090).

Module	Params	FLOPs	Time (ms)
SigLIP-2 (frozen)	400 M	12.1 G	8.2
Pseudo-RGB projection	0.01 M	0.003 G	0.1
Explicit branch	0.05 M	0.12 G	0.8
Implicit router	–	0.09 G	2.4
Adaptive filter	0.03 M	0.04 G	0.5
U-Net generator	0.28 M	1.23 G	4.3
Stage III re-encode	400 M	12.1 G	8.2
Soft-Jaccard verifier	–	0.06 G	1.0
Non-VLM lightweight path	**0.39 M**	**1.54 G**	**9.1**
Generation-only (Stages I–II)	**400.4 M**	**13.58 G**	**16.3**
Full closed loop (Stages I–IV)	**400.4 M**	**25.74 G**	**25.5**

The *non-VLM lightweight path* row aggregates all non-VLM modules; it includes the parameter-free router and verifier, and 0.39 M is its trainable parameter count. Parameter sharing and compute sharing are distinct: the frozen encoder's parameters are counted once (Stages I and III share weights), but its two forward passes process different images (input versus generated output), so both passes are counted in the FLOPs and latency of the full closed loop.

E.6 Unit activations

Figure 7 visualises representative explicit and implicit unit activations on a WV-III scene. Explicit units form compact, high-confidence supports over object-like primitives. Implicit units span elongated or compound layouts such as runway–apron corridors and harbour–dock–vessel relations. The median support area is 0.6% for s_k=1, 4.3% for s_k=2, and 24.7% for s_k=3. Each row corresponds to a single unit at one of the three scale tags

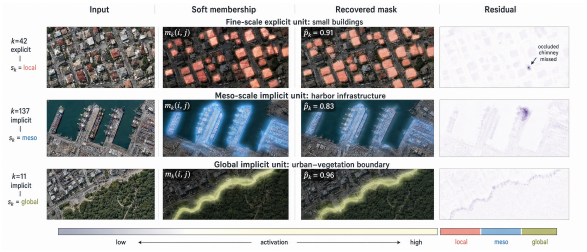


Figure 7: Explicit and implicit unit activations from Stage I on a WV-III scene. Each row shows the membership map of a single unit overlaid on the input tile. Implicit units (right) span relational structures invisible to single-scale conditioning.

(s_k=1/2/3); the soft membership column shows the differentiable presence mass $m_k(i,j)$ and the recovered mask column shows the verifier response $\tilde{\mathbf{M}}_k$ after re-encoding the generated image. The residual column highlights pixels where the recovered mask disagrees with the input membership and is the signal that $\mathcal{L}_{\text{cov}}$ minimises during training.

F Cross-Domain Pilot: Histopathology Super-Resolution

The introduction motivates semantic collapse with whole-slide pathology, whose concept hierarchy (cell nuclei → glandular structures → tumour–stroma layout) mirrors the local/meso/global pyramid. To test whether the CERES recipe transfers beyond remote sensing, we run a deliberately small pilot on 4× histopathology super-resolution.

Setup. We extract 8,000 training and 1,000 test 256×256 tiles at 0.5 μm/px from Camelyon16 whole-slide images, downsampling to 2 μm/px inputs. Following the recipe in Sec. 3, only the K=200 centroids are re-fitted (on 20k pathology patches); *every* hyperparameter of Table 18 — τ, R, $\alpha_{1:3}$, θ_{IoU}, loss weights, warm-up schedule — is reused unchanged. Prompt templates keep their schema with domain nouns (e.g. "n_k instances of {mitotic figure}"). The independent probe is a DINOv2 linear probe fit on PCam tumour/normal and NCT-CRC-HE-100K tissue-type labels, never seen in training. A 30-query bank covers 10 cellular, 10 glandular, and 10 tissue-level queries.

Table 20: Histopathology 4× SR pilot (Camelyon16 test tiles).

Method	PSNR↑	SSIM↑	$\mathcal{C}^{\text{DINO}}$↑	R@5↑
Bicubic	27.8	0.86	0.58	0.44
SwinIR-light	31.2	0.91	0.66	0.55
Open-loop SigLIP U-Net	31.4	0.91	0.70	0.58
CERES (closed-loop)	**31.6**	**0.92**	**0.83**	**0.71**

Results. Table 20 reproduces the remote-sensing

pattern: pixel metrics are nearly saturated across learning-based methods (31.4 vs. 31.6 PSNR), yet closed-loop coverage verification lifts label-grounded coverage from 0.70 to 0.83 and mean Recall@5 from 0.58 to 0.71 ($+13$ pp). The gain concentrates on cellular-scale queries ($0.40 \rightarrow 0.58$ R@5), the pathology analogue of the small-object result of Table 7. The pilot is small and uses a single organ site, so we do not claim clinical validity; its role is to show that the closed-loop recipe, with unchanged hyperparameters and only re-fitted centroids, is not specific to satellite imagery.