GeoCo-SAVi: Geometry-Consistent Slot Attention for Explicitly Editable Object Representations

Anonymous authors
Paper under double-blind review

Abstract

Object-centric video models represent scenes with slots, yet exposed geometry can vary in meaning with appearance. In Invariant Slot Attention (ISA), explicit position and scale can disagree with the decoded center and extent; edits can yield unexpected motion or resizing, and replacing appearance can shift geometry. **GeoCo-SAVi** promotes **geometric authority** and **semantic alignment**. Its spatially equivariant, object-wise decoder makes position and scale effective commands: changing them moves or resizes the rendered support. Factual position alignment ties position to the decoded center, and normalized attention overlap discourages duplicate allocation. Appearance transplantation aligns geometry semantics across objects, so recipient geometry governs layout while donor appearance supplies shape. On Obj3D, GeoCo-SAVi matches ISA reconstruction, reduces latent-position-to-decoded-centroid error by nearly 90%, and reduces appearance-induced size variation while producing the expected translation and scale responses. On 250 MOVi-C videos, it improves reconstruction and instance grouping over both same-protocol references, and video editing over STAITUS. GeoCo-SAVi transforms explicit geometry into compositional control, making both position and scale more readable and editable. Project page: `https://GeoCo-SAVi.github.io/`.

1 Introduction

Slot-based models provide object-centric vectors for decomposition, tracking, and compositional generation. Their appeal for editing rests on a simple promise: changing an object's representation should predictably change that object. Slot Attention (Locatello et al., 2020) discovers such vectors through competitive binding, and SAVi (Kipf et al., 2022) propagates them through video. Most variants still mix appearance with image-plane geometry.

ISA (Biza et al., 2023) improves this interface by exposing an appearance vector, two-dimensional position, and scale for every slot. Its relative coordinates reduce appearance's geometric burden. Editing the exposed variables can still yield unexpected motion or resizing. Their absolute meanings may vary across appearances: position can miss the rendered center, equal scales can yield different extents, and appearance swaps at fixed geometry can move or resize objects (Fig. 1).

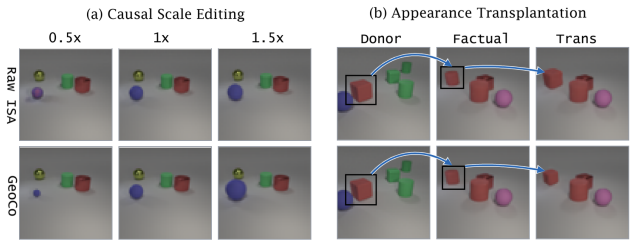

Figure 1: Raw ISA (ISA in tables) responds weakly to scale commands and can shift geometry after transplantation. GeoCo-SAVi produces predictable extent changes and preserves layout.

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.

GeoCo-SAVi addresses this gap through a geometry-control formulation that combines spatially equivariant, object-wise decoding with factual and counterfactual calibration. The decoder makes each slot's position and scale effective spatial commands: changing $\mathbf{p}_i$ translates its rendered support, while changing s_i resizes it around $\mathbf{p}_i$. Factual position alignment ties $\mathbf{p}_i$ to the decoded center, and normalized attention overlap reduces duplicate allocation.

Counterfactual supervision aligns geometry semantics across objects. Appearance transplantation combines donor appearance with recipient geometry, aligning decoded center and extent while retaining donor-dependent support structure and compactness. Appearance may remain statistically correlated with geometry; we require predictable intervention semantics.

Our contributions are:

- We make ISA's position–scale representation geometrically calibrated and explicitly editable, with consistent semantics across appearances.
- We develop a geometry-control formulation combining spatially equivariant decoding with factual and counterfactual calibration, and characterize its ideal continuous raw-logit covariance.
- We demonstrate predictable position, scale, and appearance editing on Obj3D and MOVi-C, alongside competitive reconstruction and object grouping.

2 Related Work

Object-centric image and video learning. Early generative models decompose scenes through sequential inference or mixture reconstruction (Eslami et al., 2016; Burgess et al., 2019; Engelcke et al., 2020). Slot Attention (Locatello et al., 2020) binds visual features to exchangeable slots, and SAVi (Kipf et al., 2022) extends this mechanism to conditional video tracking. Frozen self-supervised features improve grouping in realistic imagery (Seitzer et al., 2023). Video methods build on self-supervision and temporal objectives (Zadaianchuk et al., 2023; Aydemir et al., 2023; Manasyan et al., 2025), richer reasoning or synergistic learning (Seong et al., 2026; Moon & Heo, 2026), and reconstruction curricula (Moon et al., 2026). STATM (Li et al., 2025) predicts subsequent slot states through spatiotemporal attention over a memory buffer, providing a modular reasoning component for object-centric video models. These advances target discovery and tracking; our focus is the intervention semantics of exposed geometry.

Explicit factors and controllable slots. AIR, SPACE, and SCALOR use what/where factors; ISA adds a slot-centric appearance frame (Eslami et al., 2016; Lin et al., 2020b; Jiang et al., 2020; Biza et al., 2023). DISA and GOLD factor shape–texture and extrinsic–intrinsic attributes (Majellaro et al., 2025; Chen et al., 2025); STAITUS and DSSA respectively separate appearance–pose and local appearance–identity (Wei et al., 2026; Tran et al., 2026). SlotAug uses augmentation commands and CTRL-O language selection; foreground indicators refine support (Kim et al., 2023; Didolkar et al., 2025; Tian et al., 2025). We test whether $(\mathbf{p}, s)$ commands remain calibrated under novel appearance–geometry combinations.

Scale equivariance. Deep scale-space models, SESN, and DISCO are scale-equivariant hierarchies (Worrall & Welling, 2019; Sosnovik et al., 2020; 2021). SAC scales convolutional sampling within feature maps (Zhang et al., 2017). AIR and SPACE warp decoded object-local patches onto the image canvas through position and scale transformations (Eslami et al., 2016; Lin et al., 2020b). Spatial broadcast decoding combines tiled latents with coordinates (Watters et al., 2019).

3 Methodology

3.1 Preliminary: Invariant Slot Attention

For a frame $\mathbf{x} \in [0,1]^{3\times H\times W}$ on normalized grid Ω, an encoder produces tokens $\{(\mathbf{f}_n, \mathbf{u}_n)\}_{n=1}^{L}$. Our isotropic ISA (Biza et al., 2023) returns N slot representations: $\mathbf{a}_i \in \mathbb{R}^{D_a}$ is appearance, $\mathbf{p}_i \in [-1,1]^2$ is image-plane position, and $s_i > 0$ is the scalar scale:

$$\mathbf{z}_i = [\mathbf{a}_i; \mathbf{p}_i; s_i]. \tag{1}$$

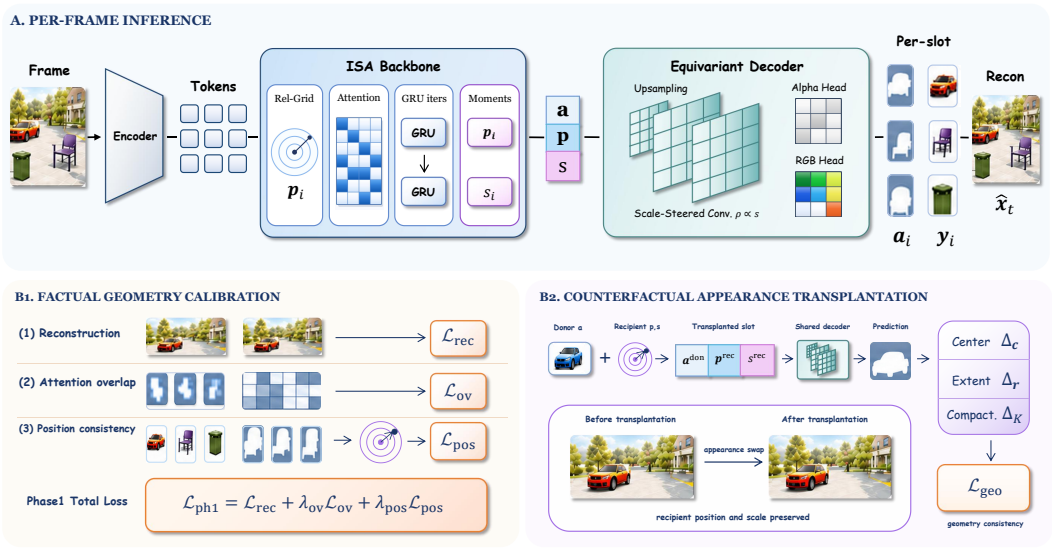

Figure 2: GeoCo-SAVi overview. *A:* ISA infers per-slot appearance and geometry; a spatially equivariant decoder renders RGB and alpha. *B1:* factual losses calibrate allocation and position. *B2:* transplants align recipient center and extent while retaining donor compactness.

The shared decoder maps each slot to RGB $\mathbf{y}_i : \Omega \to [0,1]^3$ and a raw alpha logit $\ell_i : \Omega \to \mathbb{R}$. The core contains no learned foreground, background, or empty-slot gate. Full-softmax compositing gives

$$\alpha_i(\mathbf{u}) = \mathrm{softmax}(\ell(\mathbf{u}))_i,$$
$$\hat{\mathbf{x}}(\mathbf{u}) = \sum_i \alpha_i(\mathbf{u})\mathbf{y}_i(\mathbf{u}). \tag{2}$$

Slot-centric inference. For slot i, ISA forms slot-centric token coordinates

$$\boldsymbol{\xi}_i(\mathbf{u}_n) = \frac{\mathbf{u}_n - \mathbf{p}_i}{a_s s_i + \epsilon}, \tag{3}$$

where $a_s = 5$ fixes the coordinate scale. Its encoding enters attention keys and values. If $o_i(\mathbf{u}_n)$ is slot-softmax ownership, define

$$w_i(\mathbf{u}_n) = \frac{o_i(\mathbf{u}_n)}{\sum_{m=1}^{L} o_i(\mathbf{u}_m) + \epsilon} \tag{4}$$

ISA updates appearance recurrently and reads geometry from the normalized moments

$$\mathbf{p}_i = \sum_n w_i(\mathbf{u}_n)\mathbf{u}_n, \quad s_i = \Big(\sum_n w_i(\mathbf{u}_n)\|\mathbf{u}_n - \mathbf{p}_i\|_2^2 + \epsilon\Big)^{1/2}. \tag{5}$$

Slot-indexed learnable appearance and scale initializations stabilize reconstruction, while positions are initialized IID-randomly. This relaxes strict prior exchangeability; appearance transplantation remains valid. Relative coordinates reduce appearance's geometric burden. Decoder compliance with $(\mathbf{p}_i, s_i)$ requires a commanded decoder.

Figure 2 shows the core and calibration paths; see Sec. 3.6 for implementation details.

3.2 Spatially Equivariant Decoder

The desired interface is direct: changing $\mathbf{p}_i$ should translate the rendered support, while changing s_i should resize it around $\mathbf{p}_i$. We consider two decoder implementations: scale-steered decoding on the image-plane grid, and canonical-canvas decoding followed by a position–scale warp. Both support predictable geometry-controlled editing within our formulation. We describe the scale-steered

implementation below; Appendix C.6 presents the canonical-canvas implementation and results. Scale-steered decoding uses SAC's scale-adaptive sampling rule (Zhang et al., 2017) within each slot renderer, with explicit s_i controlling its sampling footprint. For ideal continuous description, let $\mathbf{h} : \mathbb{R}^2 \to \mathbb{R}^{C_{\text{in}}}$ be a feature field and $\mathcal{K} \subset \mathbb{Z}^2$ a finite kernel support. Its action and layerwise sampling radius are

$$\begin{aligned}(\mathcal{C}_s\mathbf{h})(\mathbf{u}) &= \textstyle\sum_{\boldsymbol{\delta}\in\mathcal{K}} \mathbf{W}_{\boldsymbol{\delta}}\, \mathbf{h}(\mathbf{u} + \rho_\ell(s)\boldsymbol{\delta}) + \mathbf{b}, \\ \rho_\ell(s) &= \Delta_\ell s / s_{\text{ref}} \kappa_\ell, \\ \kappa_\ell &= \exp(\log \kappa_{\max} \tanh \psi_\ell)\,.\end{aligned} \tag{6}$$

Here Δ_ℓ is one normalized-grid pixel, $s_{\text{ref}} > 0$ is a global gauge, and κ_ℓ is an identity-initialized bounded gain. At $s = s_{\text{ref}}$ and $\kappa_\ell = 1$, the taps match ordinary stride-one convolution.

To state the requested scale response formally, let $\mathcal{T}_{\mathbf{p},k}$ rescale canvas coordinates by $k > 0$ around $\mathbf{p}$, and let $\mathcal{D}_{\mathbf{p},k}$ apply the corresponding transformation to a feature field:

$$\begin{aligned}\mathcal{T}_{\mathbf{p},k}(\mathbf{u}) &= \mathbf{p} + k(\mathbf{u} - \mathbf{p}), \\ (\mathcal{D}_{\mathbf{p},k}\mathbf{h})(\mathbf{v}) &= \mathbf{h}(\mathcal{T}_{\mathbf{p},1/k}(\mathbf{v})).\end{aligned} \tag{7}$$

Because $\rho_\ell(ks) = k\rho_\ell(s)$, the continuous operators obey

$$\mathcal{C}_{ks}\mathcal{D}_{\mathbf{p},k} = \mathcal{D}_{\mathbf{p},k}\mathcal{C}_s. \tag{8}$$

Thus commanding ks and then filtering is equivalent, in the ideal continuous setting, to filtering at s and rescaling the resulting field around the same commanded slot center.

Proposition 1 (Continuous raw-logit covariance)**.** *Let $\ell_{\mathbf{a},\mathbf{p},s}$ be the ideal continuous per-slot alpha logit generated from the slot-centric broadcast field by pointwise maps and scale-steered layers with $\rho_\ell(ks) = k\rho_\ell(s)$. With zero coordinate stabilization and no boundary truncation, for every $\mathbf{t} \in \mathbb{R}^2$ and $k > 0$,*

$$\begin{aligned}\ell_{\mathbf{a},\mathbf{p}+\mathbf{t},s}(\mathbf{u} + \mathbf{t}) &= \ell_{\mathbf{a},\mathbf{p},s}(\mathbf{u}), \\ \ell_{\mathbf{a},\mathbf{p},ks}(\mathcal{T}_{\mathbf{p},k}(\mathbf{u})) &= \ell_{\mathbf{a},\mathbf{p},s}(\mathbf{u}).\end{aligned} \tag{9}$$

Thus the ideal raw-logit path is covariant to commanded translation and scale (layerwise proof: Supplementary Material, Sec. A.1), excluding cross-slot full-softmax alpha. Finite-grid discretization, boundaries, and the bounded fixed-pixel RGB residual make equivariance approximate (Sec. 3.6).

3.3 Phase 1: Factual Geometry Calibration

Let m_i be slot i's differentiable one-foreground (OneFG) support for a detached factual background set (Eq. (19)). For any support, $\boldsymbol{\mu}$, r, A, and $K = A/(r^2+\epsilon)$ denote centroid, RMS radius, coverage, and compactness (Sec. 3.6).

We first optimize the per-frame reconstruction loss

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{c,\mathbf{u}}\left[\left(\widehat{x}_c(\mathbf{u}) - x_c(\mathbf{u})\right)^2\right]. \tag{10}$$

After warm-up, normalized overlap calibrates slot allocation:

$$\mathcal{L}_{\text{ov}} = \mathbb{E}_{i\neq j}\left[\sum\nolimits_n w_i(\mathbf{u}_n) w_j(\mathbf{u}_n)\right]. \tag{11}$$

Unit spatial mass penalizes weak duplicates (uniform slot: $1/L$) without appearance-space repulsion. The position loss aligns each valid factual OneFG centroid with its slot position:

$$\mathcal{L}_{\text{pos}} = \mathbb{E}_{i\in\mathcal{V}}\left[\overline{\mathrm{H}}_{\delta_p}(\boldsymbol{\mu}(m_i) - \mathrm{sg}(\mathbf{p}_i))\right], \tag{12}$$

Here $\overline{\mathrm{H}}_\delta$ averages coordinatewise Huber loss and $\mathcal{V}$ indexes valid supports; sg denotes stop-gradient. Only the support path receives gradients; $(\mathbf{p}_i, s_i)$ and validity are detached (piecewise form: Supplementary Material, Sec. C.2).

3.4 PHASE 2: COUNTERFACTUAL APPEARANCE TRANSPLANTATION

Factual alignment cannot constrain unseen appearance–geometry combinations. We transplant donor j into recipient i as

$$\mathbf{z}^{\mathrm{cf}}_{i\leftarrow j} = [\mathbf{a}^{\mathrm{don}}_j; \mathbf{p}^{\mathrm{rec}}_i; s^{\mathrm{rec}}_i]. \tag{13}$$

For an accepted pair, m^{rec}_i and m^{don}_j are factual supports, and $m^{\mathrm{cf}}_{i\leftarrow j}$ is decoded against the recipient's factual background. Pair construction and detached filtering are in Supplementary Material, Sec. C.7. The geometry residuals are

$$\boldsymbol{\Delta}^{ij}_c = \boldsymbol{\mu}(m^{\mathrm{cf}}_{i\leftarrow j}) - \mathrm{sg}(\mathbf{p}_i), \tag{14}$$

$$\Delta^{ij}_r = \log r(m^{\mathrm{cf}}_{i\leftarrow j}) - \mathrm{sg}(\log r(m^{\mathrm{rec}}_i)), \tag{15}$$

$$\Delta^{ij}_K = \log K(m^{\mathrm{cf}}_{i\leftarrow j}) - \mathrm{sg}(\log K(m^{\mathrm{don}}_j)). \tag{16}$$

With detached factual confidence ω_{ij}, the objective is

$$\mathcal{L}_{\mathrm{geo}} = \mathbb{E}_{(i,j)\sim\omega}\Big[\mathcal{L}_{\mathrm{ctr}}\big(\boldsymbol{\Delta}^{ij}_c\big) + \mathrm{H}_{\delta_r}\big(\Delta^{ij}_r\big) + \lambda_K \mathrm{H}_{\delta_K}\big(\Delta^{ij}_K\big)\Big]. \tag{17}$$

With factual targets and recipient commands detached, $\mathcal{L}_{\mathrm{geo}}$ updates the decoder and donor-appearance path through counterfactual alpha. Proposition 1 characterizes the geometric transformation of the donor raw-logit field under recipient commands; support consistency is encouraged by calibration and evaluated on the finite grid. $\mathcal{L}_{\mathrm{geo}}$ aligns center and RMS extent across appearances while matching donor compactness, hence $A^{\mathrm{target}}_{i\leftarrow j} = \mathrm{sg}(K^{\mathrm{don}}_j)\,\mathrm{sg}(r^{\mathrm{rec}}_i)^2$; Supplementary Material, Secs. C.3 and C.7 define the center penalty and coherence filter.

3.5 TRAINING OBJECTIVE AND CURRICULUM

The single-frame training objective is

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{ov}}\mathcal{L}_{\mathrm{ov}} + \lambda_{\mathrm{pos}}\mathcal{L}_{\mathrm{pos}} + \lambda_{\mathrm{geo}}\mathcal{L}_{\mathrm{geo}} + \lambda_{\mathrm{gate}}\mathcal{L}_{\mathrm{gate}} + \lambda_{\mathrm{tail}}\mathcal{L}_{\mathrm{tail}}. \tag{18}$$

Here $\mathcal{L}_{\mathrm{gate}}$ penalizes the mean squared terminal gate (Eq. (C.1)); the selector has a separate objective and optimizer in the Obj3D trajectory. On MOVi-C, the terminal readout, selector, and tail penalty are inactive; the results demonstrate geometry-controlled video editing without relying on these optional refinements.

We warm up reconstruction, ramp factual position and overlap, and add counterfactual geometry. Video training additionally activates slot initialization from preceding frames. The Obj3D and MOVi-C schedules, including reference-scale continuation, are in Supplementary Material, Secs. C and F, respectively. Loss weights are chosen from the observed loss and module-wise gradient magnitudes, then rounded to simple values.

3.6 IMPLEMENTATION DETAILS

Optional terminal appearance readout. On Obj3D, a bounded, zero-initialized gate refines appearance after the final readout while leaving $\mathbf{p}_i$ and s_i unchanged. The returned appearance enters both decoder branches. Supplementary Material, Sec. C.1 gives the complete residual gate and routing definition.

Discrete decoder realization. Following spatial broadcast (Watters et al., 2019), each slot is placed on a slot-centric grid; progressive bilinear synthesis then alternates with scale-steered convolutions. Every learned spatial operation on the alpha path is scale-steered. RGB additionally uses a bounded, zero-initialized fixed-pixel micro residual on Obj3D. Sampling, padding, clipping, aliasing, and cross-slot competition delimit finite-grid equivariance. Supplementary Material, Secs. C.4–C.5 give the discrete operator, heads, and topology.

Differentiable geometry support. Geometry training uses the differentiable logits of a detached factual background set $\mathcal{B}$:

$$\ell_{\mathcal{B}}(\mathbf{u}) = \mathrm{logsumexp}_{j\in\mathcal{B}}\,\ell_j(\mathbf{u}), \quad m_i(\mathbf{u}) = \sigma(\gamma[\ell_i(\mathbf{u}) - \ell_{\mathcal{B}}(\mathbf{u})]), \quad \gamma = 2. \tag{19}$$

This sharpened OneFG mask removes other foreground competitors. Invisible geometry remains outside the proxy. For nonnegative m, define

$$M_m = \sum_{\mathbf{u}} m(\mathbf{u}), \quad \boldsymbol{\mu}(m) = (M_m + \epsilon)^{-1} \sum_{\mathbf{u}} m(\mathbf{u})\mathbf{u},$$
$$r(m)^2 = (M_m + \epsilon)^{-1} \sum_{\mathbf{u}} m(\mathbf{u}) \|\mathbf{u} - \boldsymbol{\mu}(m)\|_2^2 + \epsilon, \tag{20}$$
$$A(m) = \mathbb{E}_{\mathbf{u}}[m(\mathbf{u})], \quad K(m) = (r(m)^2 + \epsilon)^{-1} A(m).$$

Under uniform scaling, r is first-order, A second-order, and K invariant (Supplementary Material, Sec. A.2). OneFG provides the differentiable training proxy; the main editing metrics use hard full-softmax ownership.

Optional selective alpha ownership. A jointly trained selector refines foreground/background ownership using detached internal evidence. Its operating point is fixed on a calibration panel from the training split. Supplementary Material, Secs. A.3–B give the ownership transfer rule, training-only teacher, and selector details.

Video slot initialization. For video inference, a STATM-style (Li et al., 2025) Transformer initializes slots from preceding slot states, followed by current-frame ISA refinement. It is trained through frame reconstruction; its residual update is detailed in Supplementary Material, Sec. C.9.

4 Experiments

We evaluate three questions: (1) whether GeoCo-SAVi preserves factual reconstruction and decomposition quality; (2) whether position and scale behave as commands; and (3) whether their meaning remains stable across appearances. Controlled single-frame Obj3D (Lin et al., 2020a) isolates these effects, while MOVi-C (Greff et al., 2022) tests them on complex video objects.

4.1 Experimental Setup and Evaluation Design

Shared protocol and three support roles. Training uses one NVIDIA A800 80GB GPU. *Training support* consists solely of model-produced decoder/attention supports, including differentiable OneFG (Eq. (19)). *Reference support* supplies external centers and pre-edit identities: model-independent RGB pseudo-labels on Obj3D and official visible GT masks on MOVi-C. *Measurement support* is the decoded mask obtained by argmax over full-softmax alpha. Editing F-scores measure adherence to prescribed position, scale, and appearance-transplantation commands; reference-based metrics assess alignment with scene geometry. All-slot evaluations include valid non-background objects without cross-method matching. Geometric response metrics use objects within the canvas to avoid clipping-induced changes in center and size; F1 uses valid edits without boundary filtering. Results without boundary filtering appear in Tables 9–10 and 19–20.

With decoded center $\hat{c}$ and reference center $c^\star$, $E_{p\to\hat{c}}$, $E_{p\to c^\star}$, and $E_{\hat{c}\to c^\star}$ denote mean L2 distances on the $[-1,1]^2$ grid. O_{attn} denotes normalized attention overlap. At fixed geometry $g = (p, s)$, $\sigma_{r|g}$ and $\sigma_{A|g}$ measure population radius and coverage variation across appearances. Radius uses normalized-grid RMS units, and coverage is the area fraction. For an edited mask M_e, the target M_t is obtained by applying the commanded transformation to the corresponding factual decoded mask. Editing F1 is $2|M_e \cap M_t|/(|M_e| + |M_t|)$. Position, scale, and appearance-transfer scores are $F_{\text{pos}}, F_{\text{scl}}, F_{\text{app}}$, respectively. All-slot empty edited outputs receive zero F1; center-based errors use outputs with defined centers.

Controlled Obj3D. Obj3D contains 64×64 scenes with moving colored solids. ISA, GeoCo-SAVi, and same-protocol SlotAug (Kim et al., 2023) use six 64-dimensional appearance states and train from scratch for 70k updates with matched capacity and sample exposure. Isotropic ISA uses reconstruction alone; GeoCo-SAVi follows Sec. 3.5; SlotAug retains its published objective and commands, with command normalization delayed to a later hidden layer. PSNR averages four inference initializations. Table 1 reports reconstruction, position alignment, fixed-geometry stability, and editing responses. External position errors cover all 2,600 windows, with 9,255 shared RGB pseudo-label objects.

Table 1: Single-frame Obj3D results. PSNR and O_{attn} cover the full validation set; $E_{p\to\hat{c}}$ and editing use the metric-specific scopes in Sec. 4. Reference errors use 9,255 shared RGB pseudo-label objects. Spreads fix $s = 0.2$, with sample counts in the supplement; editing F1 is in percent. Best values are bold.

FACTUAL RECONSTRUCTION AND GEOMETRY

Method	PSNR↑	$E_{p\to\hat{c}}\downarrow$	$E_{p\to c^\star}\downarrow$	$E_{\hat{c}\to c^\star}\downarrow$	$O_{\text{attn}}\downarrow$	$\sigma_{r\|g}\downarrow$	$\sigma_{A\|g}\downarrow$
ISA	44.145	0.1049	0.1023	0.0420	0.0148	0.0440	0.0183
SlotAug	43.216	0.0476	0.0579	**0.0358**	0.0124	0.0520	0.0226
GeoCo-SAVi	**44.210**	**0.0107**	**0.0436**	0.0391	**0.0007**	**0.0211**	**0.0086**

COUNTERFACTUAL EDITING CAPABILITY

Method	$F_{\text{pos}}\uparrow$	$F_{\text{scl}}\uparrow$	$F_{\text{app}}\uparrow$	$E_{\Delta p}\downarrow$	$\beta_r\to 1$	$\beta_A\to 2$	$D_{\text{app}}\downarrow$
ISA	86.35	62.84	75.73	0.0091	0.0939	-0.0092	0.0442
SlotAug	90.78	82.19	78.20	0.0060	0.5412	1.0575	0.0321
GeoCo-SAVi	**93.29**	**92.72**	**91.09**	**0.0051**	**1.0339**	**2.0225**	**0.0124**

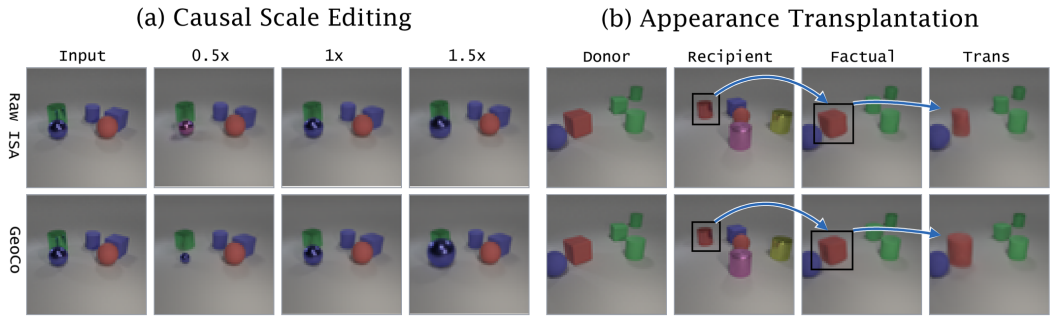

Figure 3: Counterfactual control on complementary Obj3D examples. GeoCo-SAVi produces a predictable extent response to scale commands (*left*) and places donor appearance at recipient geometry (*right*); Raw ISA is comparatively scale-insensitive and its transplanted support drifts.

MOVi-C. We use all 24 frames of the official 250-video validation set. All methods reconstruct 64×64 RGB from frozen DINOv2 features (Oquab et al., 2024). GeoCo-SAVi uses 11 slots, 64-dimensional appearance, and a 100k curriculum: two-frame clips through 70k, then four-frame clips with a zero-residual temporal initializer. Same-protocol STATM-SAVi (Li et al., 2025) and STAITUS (Wei et al., 2026) match capacity, updates, sample exposure, and common settings. STATM-SAVi uses a standard Slot Attention core with the published slot-wise spatial and temporal Transformer. STAITUS retains its temporal appearance predictor and active-set gate; baseline configurations are detailed in Supplementary Sec. F. Factual metrics cover all 250 videos. External position evaluation uses all 1,543 GT tracks with fixed pre-edit association. Complete-video editing uses 238 shared recipient/donor video pairs; geometric response diagnostics use all eligible slots.

4.2 FACTUAL RECONSTRUCTION AND DECOMPOSITION

Table 2: MOVi-C segmentation on 250 equally weighted validation videos. Best values are bold.

Method	**FG-ARI**	**ARI**	**mIoU**	**mBO**
STAITUS	0.1550	0.1114	0.1242	0.1307
STATM-SAVi	0.1619	0.4052	0.2035	0.2085
GeoCo-SAVi	**0.1935**	**0.4610**	**0.2451**	**0.2507**

PSNR (dB) is $10\log_{10}(1/\text{MSE})$ per 24-frame video for RGB in $[0, 1]$. MOVi-C full-softmax alpha is resized to GT resolution before slotwise argmax. FG-ARI, ARI, Hungarian mIoU, and mBO use all 24 frames per video, averaged equally over 250 videos. Higher values indicate better object segmentation quality on the evaluated videos.

GeoCo-SAVi matches ISA reconstruction with about 95.5% lower attention overlap (Table 1). On MOVi-C, it leads both references in PSNR and all four segmentation metrics (Tables 2 and 3).

4.3 Commanded Position and Scale

Table 3: MOVi-C reconstruction, geometry, and editing. Position errors use 1,543 GT tracks with fixed pre-edit association; PSNR and O_{attn} use all 250 videos; spreads fix $s = 0.2$, with sample counts in the supplement. F1 (%) pools all 24 frames per video and averages 238 shared recipient/donor pairs. Other response metrics use objects within the canvas. N/A denotes an unsupported geometry interface; best values are bold.

Factual Reconstruction and Geometry

Method	PSNR↑	$E_{p\to\hat{c}}\downarrow$	$E_{p\to c^\star}\downarrow$	$E_{\hat{c}\to c^\star}\downarrow$	$O_{\text{attn}}\downarrow$	$\sigma_{r\mid g}\downarrow$	$\sigma_{A\mid g}\downarrow$
STATM-SAVi	21.314	N/A	N/A	N/A	0.0045	N/A	N/A
STAITUS	19.960	0.0868	0.3417	0.3128	**0.0008**	0.1756	0.1439
GeoCo-SAVi	**22.808**	**0.0276**	**0.2403**	**0.2341**	0.0011	**0.0503**	**0.0191**

Video Editing and Geometry Response

Method	$F_{\text{pos}}\uparrow$	$F_{\text{scl}}\uparrow$	$F_{\text{app}}\uparrow$	$E_{\Delta p}\downarrow$	$\beta_r\to 1$	$\beta_A\to 2$	$D_{\text{app}}\downarrow$
STATM-SAVi	N/A	N/A	N/A	N/A	N/A	N/A	N/A
STAITUS	75.34	72.14	45.15	0.0174	0.7322	1.2122	0.0585
GeoCo-SAVi	**86.51**	**84.93**	**74.01**	**0.0067**	**0.8892**	**1.6178**	**0.0334**

Position. GeoCo-SAVi improves latent-to-decoded and latent-to-reference position alignment on both datasets (Tables 1 and 3). On Obj3D, GeoCo-SAVi's decoded centers are closer to the RGB pseudo-label centers than ISA's, while SlotAug has the lowest decoded-center reference error. Translation error $E_{\Delta p}$ measures commanded-versus-observed center displacement error, normalized by the image diagonal. Video F1 edits the same slot through all 24 frames, sums intersections and mask areas over time before applying the F1 ratio, and averages videos and commands equally. Targets follow the prescribed transformations (Sec. 4). STATM-SAVi lacks an explicit geometry-control interface.

Scale. We use factors 0.5, 0.75, 1, 1.25, and 1.5. For $x \in \{r, A\}$, the fitted log-response slope β_x has ideal target 1 for radius and 2 for coverage; editing F1 excludes the identity command. GeoCo-SAVi improves scale-edit F1 on both datasets. Its Obj3D response closely follows the ideal scaling law; MOVi-C retains finite-grid and scene-dependent deviations (Tables 1 and 3; Fig. 3).

4.4 Appearance-Consistent Geometry

GeoCo-SAVi reduces appearance-transplant drift on both datasets with recipient geometry fixed. D_{app} measures decoded-center displacement on the normalized grid. At fixed g with $s = 0.2$, radius and coverage spreads are also lower, indicating more consistent extent and area across appearances (Tables 1 and 3). Complete-video appearance-transfer F1 is 74.01% (STAITUS: 45.15%). These results and Fig. 3 support more consistent geometric commands across donor appearances.

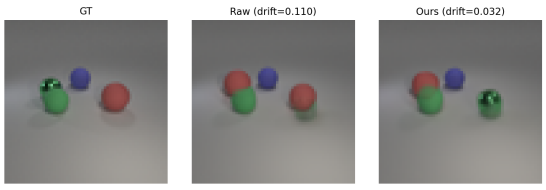

Figure 4: Appearance swap.

The appearance-swap comparison in Fig. 4 shows that factual calibration reduces appearance-swap centroid drift. Raw ISA displaces the swapped object and leaves ghost artifacts; the calibrated diagnostic preserves its recipient location. This 40k diagnostic follows a separate trajectory using reconstruction, position, and normalized-overlap losses.

STAITUS's O_{attn} is computed only on active slots, whereas the other methods include empty slots with uniform attention; this convention favors lower STAITUS values. Scopes and object-matched editing results with reference-mask-based RGB targets appear in Secs. E–F (E.6 and F.7).

4.5 Ablation Study

Seven controls assess the roles of decoder design, factual and counterfactual calibration, and optional refinements.

Table 4: Obj3D component comparisons. Full denotes GeoCo-SAVi. Response metrics use the same within-canvas scope as Table 1.

Diagnostic	Ablated	Full
Conventional decoder		
$\beta_r \to 1$	0.7242	**1.0339**
$\beta_A \to 2$	1.3929	**2.0225**
Without $\mathcal{L}_{\text{ov}}$		
$O_{\text{attn}} \downarrow$	0.0077	**0.0007**
Without $\mathcal{L}_{\text{pos}}$		
$E_{p\to\hat{c}} \downarrow$	0.0175	**0.0107**
$E_{p\to c^\star} \downarrow$	0.0536	**0.0436**
Without $\mathcal{L}_{\text{geo}}$		
$D_{\text{app}} \downarrow$	0.0305	**0.0124**
$\sigma_{r\|g} \downarrow$	0.0256	**0.0211**
$\sigma_{A\|g} \downarrow$	0.0098	**0.0086**
Without terminal readout		
PSNR↑	43.7399	**44.2099**
$E_{\hat{c}\to c^\star} \downarrow$	0.0487	**0.0391**
Without selector		
PSNR↑	44.2056	**44.2099**
$E_{\hat{c}\to c^\star} \downarrow$	0.0451	**0.0391**
Without $\mathcal{L}_{\text{tail}}$		
PSNR↑	**44.2995**	44.2099
$D_{\text{app}} \downarrow$	0.0215	**0.0124**

Decoder design. The conventional decoder yields radius and coverage slopes below their ideal targets, whereas the full model is closer to both. The slopes measure radius and coverage responses to scale commands.

Attention allocation. Without $\mathcal{L}_{\text{ov}}$, factual attention overlap rises from 0.0007 to 0.0077, indicating that spatial normalization discourages duplicate attention without a foreground/background gate.

Factual position calibration. Without $\mathcal{L}_{\text{pos}}$, latent-to-decoded center error rises from 0.0107 to 0.0175, while latent-to-reference error increases from 0.0436 to 0.0536. Factual calibration improves command alignment and slightly reduces reference-center error.

Counterfactual geometry. Without $\mathcal{L}_{\text{geo}}$, transfer centroid drift increases from 0.0124 to 0.0305; radius and coverage spreads increase from 0.0211/0.0086 to 0.0256/0.0098, supporting its role in stabilizing object centers and spatial extents across donor appearances.

Optional components. Terminal readout raises PSNR from 43.7399 to 44.2099 and lowers decoded-center reference error from 0.0487 to 0.0391. The selector also improves both metrics. The tail penalty trades a small PSNR decrease (44.2995 to 44.2099) for lower appearance-transplant drift (0.0215 to 0.0124).

All variants are evaluated at 70k updates using the same metric definitions and evaluation procedures as the full model. PSNR is averaged over four inference initializations. Component settings and training schedules are detailed in Supplementary Sec. E.7.

5 Conclusion

Summary. GeoCo-SAVi connects explicit position–scale representations to decoded object geometry, enabling calibrated readout and direct editing with consistent semantics across appearances. On Obj3D, it matches ISA reconstruction with better geometry and editing. On MOVi-C, it improves reconstruction and grouping over same-protocol references and video editing over STAITUS.

Limitations. Equivariance is established for continuous per-slot raw logits; discretization, clipping, and cross-slot softmax limit its finite-grid realization. Scale is isotropic, decoded support need not match physical object boundaries, and occlusion can impair geometric recovery. Video-training cost limits configuration and hyperparameter coverage.

Future work. Priorities include broader validation, explicit depth and occlusion modeling, and combining calibrated image-plane representations with camera and scene geometry to recover real-world object trajectories for world-model learning and prediction under partial visibility.

AI USAGE STATEMENT

Generative AI tools, under the authors' direction and judgment, provided support for the development and refinement of hypotheses, theoretical and conceptual formulations, mathematical claims, mathematical reasoning and proof preparation, research methodology and experimental design, method implementation, result interpretation, and translation. Additional support included literature search and summarization, manuscript editing, and preparation of code and documentation.

Across these activities, the authors guided the work, reviewed and verified the resulting AI-assisted material, and made the final decisions regarding its incorporation. The authors take responsibility for the final text, scientific claims, experimental results, code, and accompanying materials.

REPRODUCIBILITY STATEMENT

We provide the implementation of GeoCo-SAVi, dataset preparation and loading scripts for Obj3D and MOVi-C, training entry points, configurations, and the metric primitives used in this work through an anonymous code repository.[1] The released code covers the slot-attention core, spatially equivariant decoder, factual and counterfactual objectives, temporal initializer, optional refinements, and the associated optimizer and gradient-routing logic.

The main paper and supplementary material document the model architecture, theoretical formulation, training objectives and curricula, settings, evaluation metrics, sample populations, aggregation rules, denominators, and finite-grid diagnostics. Together with the released configurations and code, these materials specify the procedures used to obtain and evaluate the reported results.

ETHICS STATEMENT

This work studies object-centric representation learning and controllable video representations on the synthetic Obj3D and MOVi-C benchmarks. It involves no human participants, personal data, or sensitive attributes. As with other learned visual models, deployment beyond the evaluated synthetic setting could inherit biases from new data or enable misleading manipulation of visual content. Such uses require evaluation on the intended data, appropriate safeguards, and human oversight. The authors take responsibility for the research artifacts and their use.

REFERENCES

Görkay Aydemir, Weidi Xie, and Fatma Güney. Self-supervised object-centric learning for videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ondrej Biza, Sjoerd van Steenkiste, Mehdi S. M. Sajjadi, Gamaleldin F. Elsayed, Aravindh Mahendran, and Thomas Kipf. Invariant slot attention: Object discovery with slot-centric reference frames. In *International Conference on Machine Learning (ICML)*, 2023.

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Tonglin Chen, Yinxuan Huang, Zhimeng Shen, Jinghao Huang, Bin Li, and Xiangyang Xue. Learning global object-centric representations via disentangled slot attention. *Machine Learning*, 114 (2):40, 2025.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Aniket Didolkar, Andrii Zadaianchuk, Rabiul Awal, Maximilian Seitzer, Efstratios Gavves, and Aishwarya Agrawal. CTRL-O: Language-controllable object-centric visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29523–29533, 2025.

[1] `https://github.com/GeoCo-SAVi/GeoCo-SAVi`

Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. In *International Conference on Learning Representations (ICLR)*, 2020.

S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, et al. Kubric: A scalable dataset generator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, and Sungjin Ahn. SCALOR: Generative world models with scalable object representations. In *International Conference on Learning Representations (ICLR)*, 2020.

Jinwoo Kim, Janghyuk Choi, Jaehyun Kang, Changyeon Lee, Ho-Jin Choi, and Seon Joo Kim. Leveraging image augmentation for object manipulation: Towards interpretable controllability in object-centric learning. *arXiv preprint arXiv:2310.08929*, 2023.

Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations (ICLR)*, 2022.

Jian Li, Pu Ren, Yang Liu, and Hao Sun. Reasoning-enhanced object-centric learning for videos. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 659–670. Association for Computing Machinery, 2025. doi: 10.1145/3690624.3709168.

Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving generative imagination in object-centric world models. In *International Conference on Machine Learning (ICML)*, 2020a.

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. SPACE: Unsupervised object-oriented scene representation via spatial attention and decomposition. In *International Conference on Learning Representations (ICLR)*, 2020b.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

Riccardo Majellaro, Jonathan Collu, Aske Plaat, and Thomas M. Moerland. Explicitly disentangled representations in object-centric learning. *Transactions on Machine Learning Research*, 2025.

Anna Manasyan, Maximilian Seitzer, Filip Radovic, Georg Martius, and Andrii Zadaianchuk. Temporally consistent object-centric learning by contrasting slots. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5401–5411, 2025.

WonJun Moon and Jae-Pil Heo. Selective synergistic learning for video object-centric learning. *arXiv preprint arXiv:2606.15527*, 2026.

WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Reconstruction-guided slot curriculum: Addressing object over-fragmentation in video object-centric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25001–25010, 2026.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Hyun Seok Seong, WonJun Moon, and Jae-Pil Heo. From vicious to virtuous cycles: Synergistic representation learning for unsupervised video object-centric learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Ivan Sosnovik, Michał Szmaja, and Arnold W. M. Smeulders. Scale-equivariant steerable networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Ivan Sosnovik, Artem Moskalev, and Arnold W. M. Smeulders. DISCO: Accurate discrete scale convolutions. In *British Machine Vision Conference (BMVC)*, 2021.

Pinzhuo Tian, Shengjie Yang, Hang Yu, and Alex C. Kot. Pay attention to the foreground in object-centric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30281–30290, 2025.

Sieu Tran, Duc Nguyen, Hao Vo, Khoa Vo, and Ngan Le. Dual-state slot attention: Decoupling appearance and identity for video object-centric learning. *arXiv preprint arXiv:2606.12601*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Nicholas Watters, Loic Matthey, Christopher P. Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *arXiv preprint arXiv:1901.07017*, 2019.

Amaury Wei, Ismail Nejjar, and Olga Fink. Rethinking object-centric representations for video dynamics modeling. *arXiv preprint arXiv:2606.23436*, 2026.

Daniel E. Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 7364–7376, 2019.

Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

SUPPLEMENTARY MATERIAL

A CONTINUOUS EQUIVARIANCE AND SCALING PROOFS

A.1 SCALE-STEERED CONVOLUTION

Lemma 1 (Layer commutation)**.** *Let $\mathcal{C}_s$ be the continuous scale-steered layer in Eq.* (6)*, acting on fields over $\mathbb{R}^2$, and assume $\rho_\ell(ks) = k\rho_\ell(s)$ for every $k > 0$. Define translation by $(\mathcal{D}_{\mathbf{t}}\mathbf{h})(\mathbf{v}) = \mathbf{h}(\mathbf{v} - \mathbf{t})$. With no boundary truncation,*

$$\mathcal{C}_{ks}\mathcal{D}_{\mathbf{p},k} = \mathcal{D}_{\mathbf{p},k}\mathcal{C}_s, \qquad \mathcal{C}_s\mathcal{D}_{\mathbf{t}} = \mathcal{D}_{\mathbf{t}}\mathcal{C}_s. \tag{A.1}$$

Every pointwise affine map or nonlinearity commutes with both actions as well.

Proof. For scale covariance, evaluate the left-hand side at $\mathcal{T}_{\mathbf{p},k}(\mathbf{u})$:

$$\begin{aligned}
&(\mathcal{C}_{ks}\mathcal{D}_{\mathbf{p},k}\mathbf{h})(\mathcal{T}_{\mathbf{p},k}(\mathbf{u})) \\
&\quad= \sum_{\boldsymbol{\delta}\in\mathcal{K}} \mathbf{W}_{\boldsymbol{\delta}}(\mathcal{D}_{\mathbf{p},k}\mathbf{h})(\mathcal{T}_{\mathbf{p},k}(\mathbf{u}) + \rho_\ell(ks)\boldsymbol{\delta}) + \mathbf{b} \\
&\quad= \sum_{\boldsymbol{\delta}\in\mathcal{K}} \mathbf{W}_{\boldsymbol{\delta}}\mathbf{h}\left(\mathbf{u} + \frac{\rho_\ell(ks)}{k}\boldsymbol{\delta}\right) + \mathbf{b} \\
&\quad= \sum_{\boldsymbol{\delta}\in\mathcal{K}} \mathbf{W}_{\boldsymbol{\delta}}\mathbf{h}(\mathbf{u} + \rho_\ell(s)\boldsymbol{\delta}) + \mathbf{b} = (\mathcal{C}_s\mathbf{h})(\mathbf{u}).
\end{aligned} \tag{A.2}$$

The third line applies $\mathcal{T}_{\mathbf{p},1/k}$ to the sampling coordinate, and the fourth uses $\rho_\ell(ks) = k\rho_\ell(s)$. By the definition of $\mathcal{D}_{\mathbf{p},k}$, the final expression equals $(\mathcal{D}_{\mathbf{p},k}\mathcal{C}_s\mathbf{h})(\mathcal{T}_{\mathbf{p},k}(\mathbf{u}))$. Translation follows from

$$(\mathcal{C}_s\mathcal{D}_{\mathbf{t}}\mathbf{h})(\mathbf{u} + \mathbf{t}) = \sum_{\boldsymbol{\delta}\in\mathcal{K}} \mathbf{W}_{\boldsymbol{\delta}}\mathbf{h}(\mathbf{u} + \rho_\ell(s)\boldsymbol{\delta}) + \mathbf{b} = (\mathcal{C}_s\mathbf{h})(\mathbf{u}). \tag{A.3}$$

Finally, for a pointwise map ϕ and spatial transformation $\mathcal{T}$ inducing $\mathcal{D}$, $\phi(\mathcal{D}\mathbf{h})(\mathbf{v}) = \phi(\mathbf{h}(\mathcal{T}^{-1}\mathbf{v})) = (\mathcal{D}\phi(\mathbf{h}))(\mathbf{v})$. □

Theorem 1 (Per-slot raw-logit covariance)**.** *Set the coordinate stabilizer to zero. Let the initial per-slot field be*

$$\mathbf{h}^{(0)}_{\mathbf{a},\mathbf{p},s}(\mathbf{u}) = \mathrm{Broadcast}(\mathbf{a}) + \mathrm{PE}\left(\frac{\mathbf{u} - \mathbf{p}}{a_s s}\right), \tag{A.4}$$

and let $\ell_{\mathbf{a},\mathbf{p},s}$ be obtained from this field by any finite composition of pointwise maps and scale-steered layers satisfying the assumptions of Lemma 1, followed by a pointwise alpha-logit head. Then

$$\begin{aligned}
\ell_{\mathbf{a},\mathbf{p}+\mathbf{t},s}(\mathbf{u} + \mathbf{t}) &= \ell_{\mathbf{a},\mathbf{p},s}(\mathbf{u}), \\
\ell_{\mathbf{a},\mathbf{p},ks}(\mathcal{T}_{\mathbf{p},k}(\mathbf{u})) &= \ell_{\mathbf{a},\mathbf{p},s}(\mathbf{u})
\end{aligned} \tag{A.5}$$

for every $\mathbf{t} \in \mathbb{R}^2$ and $k > 0$.

Proof. The initial field is translation covariant because

$$\frac{(\mathbf{u} + \mathbf{t}) - (\mathbf{p} + \mathbf{t})}{a_s s} = \frac{\mathbf{u} - \mathbf{p}}{a_s s},$$

and scale covariant because

$$\frac{\mathcal{T}_{\mathbf{p},k}(\mathbf{u}) - \mathbf{p}}{a_s ks} = \frac{\mathbf{u} - \mathbf{p}}{a_s s}.$$

Thus Eq. (A.5) holds at layer zero. Lemma 1 preserves the corresponding identity after each scale-steered or pointwise layer. Induction over the finite decoder path, including the pointwise alpha head, proves the result and hence Proposition 1 in the main paper. □

Remark 1 (Scope)**.** *The theorem is exact only for the ideal continuous per-slot raw-logit operator. Finite-grid interpolation, padding, clipping, nonzero numerical stabilization, and boundary truncation are excluded. If all slot logits, including background logits, undergo the same global action, pointwise softmax also covaries. Under an intervention that changes only one slot, the other logits remain fixed and cross-slot full-softmax competition generally prevents exact alpha equivariance.*

A.2 Support Moments

With $\epsilon = 0$ and finite, positive mass and radius, assume an isolated continuous support obeys

$$m_{ks}(\mathcal{T}_{\mathbf{p},k}(\mathbf{u})) = m_s(\mathbf{u}). \tag{A.6}$$

The substitution $\mathbf{v} = \mathcal{T}_{\mathbf{p},k}(\mathbf{u})$ has area element $d\mathbf{v} = k^2 d\mathbf{u}$. It follows that

$$\boldsymbol{\mu}(ks) = \mathbf{p} + k[\boldsymbol{\mu}(s) - \mathbf{p}], \qquad r(ks) = kr(s), \tag{A.7}$$

and $A(ks) = k^2 A(s)$ whenever the support is not clipped by the canvas. Consequently,

$$K(ks) = \frac{k^2 A(s)}{k^2 r(s)^2} = K(s). \tag{A.8}$$

Thus Eq. (17) jointly aligns radius and compactness.

A.3 Minimal Owner-to-Background Transfer

For an accepted pixel, let $d_i = \ell_i - \ell_{\mathrm{bg}}$ and $d_{\mathrm{alt}} = \ell_{\mathrm{alt}} - \ell_{\mathrm{bg}}$ denote the owner and strongest alternative foreground gaps. The minimum symmetric transfer needed for background to win by margin m_{bg} is

$$\Delta_{\mathrm{req}}(x) = \max\{0,\, (d_i + m_{\mathrm{bg}})/2,\, d_{\mathrm{alt}} + m_{\mathrm{bg}}\}. \tag{A.9}$$

Its capped update is

$$\Delta_{\mathrm{tr}} = \min(\Delta_{\mathrm{req}}, \Delta_{\max}), \quad \ell_i' = \ell_i - \Delta_{\mathrm{tr}}, \quad \ell_{\mathrm{bg}}' = \ell_{\mathrm{bg}} + \Delta_{\mathrm{tr}}. \tag{A.10}$$

The action is zero outside accepted pixels. Its threshold and cap are fixed on a calibration panel drawn only from the training split.

Under the symmetric action $\ell_i' = \ell_i - \Delta_{\mathrm{tr}}$ and $\ell_{\mathrm{bg}}' = \ell_{\mathrm{bg}} + \Delta_{\mathrm{tr}}$, background beats the current owner by margin m_{bg} iff

$$\begin{aligned} &\ell_{\mathrm{bg}} + \Delta_{\mathrm{tr}} \geq \ell_i - \Delta_{\mathrm{tr}} + m_{\mathrm{bg}}, \\ &\qquad \Longleftrightarrow \quad \Delta_{\mathrm{tr}} \geq \frac{\ell_i - \ell_{\mathrm{bg}} + m_{\mathrm{bg}}}{2}. \end{aligned} \tag{A.11}$$

It beats the strongest untouched foreground competitor iff

$$\begin{aligned} &\ell_{\mathrm{bg}} + \Delta_{\mathrm{tr}} \geq \ell_{\mathrm{alt}} + m_{\mathrm{bg}}, \\ &\qquad \Longleftrightarrow \quad \Delta_{\mathrm{tr}} \geq \ell_{\mathrm{alt}} - \ell_{\mathrm{bg}} + m_{\mathrm{bg}}. \end{aligned} \tag{A.12}$$

Together with nonnegativity, Eq. (A.9) is the smallest symmetric transfer satisfying both margins. Equation (A.10) applies its capped value and satisfies them whenever the cap does not bind.

B Teacher–Selector Details

B.1 Background and Deployment Domain

Let $\partial\Omega$ be a narrow image border. The background slot is selected from detached alpha by

$$i_{\mathrm{bg}} = \arg\max_i \left[\frac{1}{|\partial\Omega|} \sum_{x \in \partial\Omega} \alpha_i(x) + \lambda_{\mathrm{mass}} \frac{1}{|\Omega|} \sum_{x \in \Omega} \alpha_i(x) \right]. \tag{B.1}$$

The deployment domain contains only foreground-owned pixels whose owner has sufficient hard area and peak alpha and whose required transfer does not exceed a labeling limit. These conditions are detached and are identical when producing teacher labels and applying the trained selector.

B.2 High-Precision Tri-State Evidence

For an owner i and background i_{bg}, define the individual renderer errors

$$E_i(x) = \frac{1}{3}\|\mathbf{y}_i(x) - \mathbf{x}(x)\|_2^2, \tag{B.2}$$

$$E_{\mathrm{bg}}(x) = \frac{1}{3}\|\mathbf{y}_{i_{\mathrm{bg}}}(x) - \mathbf{x}(x)\|_2^2. \tag{B.3}$$

Let $E_{\mathrm{cur}}(x)$ and $E_{\mathrm{act}}(x)$ be the composite errors before and after the analytic transfer. We use

$$G_{\mathrm{act}}(x) = \frac{E_{\mathrm{cur}}(x) - E_{\mathrm{act}}(x)}{E_{\mathrm{cur}}(x) + E_{\mathrm{act}}(x) + \epsilon}, \tag{B.4}$$

$$G_{\mathrm{ind}}(x) = \frac{E_i(x) - E_{\mathrm{bg}}(x)}{E_i(x) + E_{\mathrm{bg}}(x) + \epsilon}, \tag{B.5}$$

where positive G_{ind} favors background. Pixels for which the action is demonstrably harmful or the foreground renderer is substantially better form the protected set,

$$c_-(x) = \mathbb{I}[G_{\mathrm{act}}(x) \le \theta_{\mathrm{act}}^-] \vee \mathbb{I}[G_{\mathrm{ind}}(x) \le \theta_{\mathrm{ind}}^-]. \tag{B.6}$$

The positive teacher combines two complementary high-precision branches. The first, c_{act}^+, requires a positive counterfactual action gain and conservative attention/alpha-density checks. The second, c_{local}^+, uses a local target-image background prototype to break the feedback loop in which a pixel captured by foreground never trains the background renderer. Let $i(u) = \arg\max_k \alpha_k(u)$ denote the decoded hard owner at pixel u. For a neighborhood $\mathcal{N}(x)$,

$$\boldsymbol{\mu}_{\mathrm{local}}(x) = \frac{\sum_{u\in\mathcal{N}(x)} \mathbb{I}[i(u) = i_{\mathrm{bg}}]\mathbf{x}(u)}{\sum_{u\in\mathcal{N}(x)} \mathbb{I}[i(u) = i_{\mathrm{bg}}] + \epsilon}, \tag{B.7}$$

$$E_{\mathrm{local}}(x) = \frac{1}{3}\|\mathbf{x}(x) - \boldsymbol{\mu}_{\mathrm{local}}(x)\|_2^2. \tag{B.8}$$

The branch requires enough background-owned neighbors, a low normalized local error, and a feasible transfer magnitude. The raw positive confidence and conflict-resolved label are

$$\begin{aligned} c_+(x) &= \max\{c_{\mathrm{act}}^+(x), c_{\mathrm{local}}^+(x)\}, \\ \widetilde{c}_+(x) &= c_+(x)[1 - c_-(x)]. \end{aligned} \tag{B.9}$$

B.3 Selector Features and Ranking Loss

The raw pointwise evidence has 139 dimensions: eight ownership/logit/density scalars, a 64-dimensional owner feature, a 64-dimensional owner–background feature difference, and a three-dimensional equivariant RGB prediction difference. One shared 1×1 projection maps the owner feature to eight dimensions and, with the same weights and canceled bias, maps the owner–background difference to eight dimensions, giving a 27-dimensional selector input. With SiLU activation ϕ,

$$\begin{aligned} \mathbf{r}_1(x) &= \phi(\mathbf{W}_1\mathbf{f}(x) + \mathbf{b}_1), \\ \mathbf{r}_2(x) &= \phi(\mathbf{W}_2\mathbf{r}_1(x) + \mathbf{b}_2), \\ \pi_{\mathrm{sel}}(x) &= \sigma(\mathbf{w}_3^\top \mathbf{r}_2(x) + b_3). \end{aligned} \tag{B.10}$$

Hidden widths 32 and 16 yield 1,961 trainable parameters.

Let $\nu(x)$ be the selector logit. Class losses are macro-averaged over nonempty scene–owner groups. For groups with both positive and protected pixels, soft extrema

$$\nu_+^{\min} = -T_{\mathrm{rank}} \log \frac{\sum_x \widetilde{c}_+(x) e^{-\nu(x)/T_{\mathrm{rank}}}}{\sum_x \widetilde{c}_+(x) + \epsilon}, \tag{B.11}$$

$$\nu_-^{\max} = T_{\mathrm{rank}} \log \frac{\sum_x c_-(x) e^{\nu(x)/T_{\mathrm{rank}}}}{\sum_x c_-(x) + \epsilon} \tag{B.12}$$

define

$$\mathcal{L}_{\text{rank}} = T_{\text{rank}} \operatorname{softplus}\left(\frac{m_{\text{rank}} + \nu_{-}^{\max} - \nu_{+}^{\min}}{T_{\text{rank}}}\right). \tag{B.13}$$

Ranking averages only these groups, with zero loss if none qualify. Unknown pixels contribute zero selector gradient. With group-balanced positive and protected logistic losses $\mathcal{L}_{+}$ and $\mathcal{L}_{-}$, the complete objective is

$$\mathcal{L}_{\text{sel}} = \mathcal{L}_{+} + \lambda_{-}\mathcal{L}_{-} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}. \tag{B.14}$$

C Implementation and Curriculum Details

All experiments reported in the main paper use one NVIDIA A800 80GB GPU.

C.1 Terminal Appearance Readout

ISA's final value aggregate is converted by the shared recurrent update into an appearance proposal $\widetilde{\mathbf{a}}_i$. The bounded residual readout is

$$\begin{aligned} \mathbf{a}_i^{\text{out}} &= \mathbf{a}_i + \mathbf{g}_i \odot (\widetilde{\mathbf{a}}_i - \mathbf{a}_i), \\ \mathbf{g}_i &= g_{\text{read}} \tanh(\mathbf{W}_g \operatorname{LN}[\mathbf{a}_i; \widetilde{\mathbf{a}}_i - \mathbf{a}_i] + \mathbf{b}_g). \end{aligned} \tag{C.1}$$

The bounded gate $\mathbf{g}_i$ is initialized to zero, giving an identity appearance update. The returned $\mathbf{a}_i^{\text{out}}$ enters both decoder branches and is subsequently abbreviated as $\mathbf{a}_i$. It refines the appearance representation used for RGB reconstruction without directly updating $\mathbf{p}_i$ or s_i.

C.2 Huber Loss

For a scalar residual e and threshold $\delta > 0$, the notation used in the main paper is

$$\mathrm{H}_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le \delta, \\ \delta\left(|e| - \frac{1}{2}\delta\right), & |e| > \delta. \end{cases} \tag{C.2}$$

For $\mathbf{e} \in \mathbb{R}^d$, the overlined form denotes the coordinate mean,

$$\overline{\mathrm{H}}_\delta(\mathbf{e}) = \frac{1}{d}\sum_{q=1}^{d} \mathrm{H}_\delta(e_q). \tag{C.3}$$

C.3 Center and Tail Penalties

For the center residual $\boldsymbol{\Delta} = \boldsymbol{\Delta}_c^{ij}$ in Eq. (17), recipient scale s, and global training step t, define

$$\mathcal{L}_{\text{ctr}}(\boldsymbol{\Delta}; s, t) = \begin{cases} \frac{1}{2}\sum_{d\in\{x,y\}} \mathrm{H}_{\delta_c}(\Delta_d), & \text{coordinate-wise}, \\ \mathrm{H}_{\delta_c}\left(\frac{\|\boldsymbol{\Delta}\|_2}{s+\epsilon}\right), & \text{scale-normalized}. \end{cases} \tag{C.4}$$

MOVi-C uses the coordinate-wise form throughout geometry training. Obj3D uses the coordinate-wise form before step 50,000 and the scale-normalized form thereafter, with $\delta_c = 0.05$. Recipient commands remain detached. The main text abbreviates this penalty as $\mathcal{L}_{\text{ctr}}(\boldsymbol{\Delta}_c^{ij})$, suppressing its scale and training-step dependence; the factual position loss is unchanged.

The tail penalty acts on gamma-2 OneFG support $q(x) = \operatorname{sigmoid}(2z(x))$, where z is the target logit minus the aggregated background logit. For pixel grid Ω, let

$$T(q) = -\frac{1}{|\Omega|}\sum_{x\in\Omega} \operatorname{sg}[\mathbf{1}\{q(x) < \tau\}] \log(1 - q(x)), \qquad \tau = 0.25. \tag{C.5}$$

The implementation evaluates $-\log(1-q)$ stably as $\operatorname{softplus}(2z)$. The detached threshold selects low-confidence pixels only; no positive-target term sharpens the remaining pixels. We combine factual and counterfactual penalties as

$$\mathcal{L}_{\text{tail}} = \tfrac{1}{2}\mathbb{E}_{\text{FG}}[T(q^{\text{fact}})] + \tfrac{1}{2}\mathbb{E}_{(i,j)\sim\omega}[T(q^{\text{cf}}_{i\leftarrow j})]. \tag{C.6}$$

The factual mean uses accepted foreground slots with a nonempty background set; the counterfactual mean uses the detached pair-confidence weights of the geometry objective and the recipient's factual background logits. Spatial means are taken over the full canvas. Empty eligible populations contribute zero. Obj3D activates $\lambda_{\text{tail}} = 10^{-3}$ at step 60,000 without an additional ramp; MOVi-C uses $\lambda_{\text{tail}} = 0$.

C.4 Discrete Scale-Steered Convolution

Let $\mathbf{H} \in \mathbb{R}^{C_{\text{in}} \times H_\ell \times W_\ell}$ be a discrete feature map and let $\mathbf{q} \in \mathbb{Z}^2$ index an output pixel. In feature-pixel units, the commanded tap dilation is

$$d_\ell(s) = \frac{s}{s_{\text{ref}}} \kappa_\ell. \tag{C.7}$$

The implemented scale-steered cross-correlation is

$$(\widehat{\mathcal{C}}_s \mathbf{H})_o[\mathbf{q}] = b_o + \sum_{c=1}^{C_{\text{in}}} \sum_{\boldsymbol{\delta} \in \mathcal{K}} W_{o,c,\boldsymbol{\delta}} \, \text{Bilin}(\mathbf{H}_c, \mathbf{q} + d_\ell(s)\boldsymbol{\delta}) \, . \tag{C.8}$$

Here $\mathbf{W} = [W_{o,c,\boldsymbol{\delta}}]$ is the learned kernel tensor and Bilin is bilinear sampling on the current feature grid. Pixel centers are converted to normalized $[-1, 1]^2$ coordinates with aligned corner centers; samples outside the finite canvas use zero padding. When $s = s_{\text{ref}}$ and $\kappa_\ell = 1$, all taps are integer-aligned and Eq. (C.8) reduces to the corresponding ordinary stride-one convolution up to floating-point interpolation error.

Each rendered slot retains its own scalar s. The implementation constructs all $|\mathcal{K}|$ analytic tap locations directly from Eq. (C.7); no network predicts deformable offsets. The taps are evaluated jointly by vectorized bilinear sampling, followed by the same channel contraction as a conventional convolution. This implementation and the reference per-tap implementation are algebraically identical, and gradients propagate through the sampler to the input field, scale command, and bounded layer gain.

C.5 Decoder Topology

Following spatial broadcast (Watters et al., 2019), slot i induces

$$\mathbf{h}_i^{(0)}(\mathbf{u}) = \text{Broadcast}(\mathbf{a}_i) + \text{PE}\left(\frac{\mathbf{u} - \mathbf{p}_i}{a_s s_i + \epsilon}\right). \tag{C.9}$$

A learned pointwise embedding PE is injected once, after which progressive bilinear synthesis alternates with bilinearly sampled scale-steered convolutions. The pointwise alpha head is

$$\ell_i(\mathbf{u}) = \mathbf{w}_\alpha^\top \mathbf{h}_i^{\text{eq}}(\mathbf{u}) + b_\alpha, \tag{C.10}$$

so every learned spatial operation on its path is scale-steered. RGB uses a pointwise fusion of $\mathbf{h}_i^{\text{eq}}$ with broadcast appearance to form $\mathbf{h}_i^{\text{rgb}}$, followed by

$$\mathbf{y}_i(\mathbf{u}) = \sigma\Big(\mathbf{y}_i^{\text{eq}}(\mathbf{u}) + \eta_{\max} \tanh\Big[\mathcal{C}_{3\times3}^{\text{micro}} \mathbf{h}_i^{\text{rgb}}(\mathbf{u})\Big]\Big) \, , \tag{C.11}$$

where the fixed-pixel micro convolution is zero-initialized and the residual is bounded by $\eta_{\max}$. We enable it for Obj3D RGB; being bounded, it refines appearance while leaving the alpha and geometry paths untouched. Here $\mathbf{h}_i^{\text{eq}}$ is the final equivariant feature field and $\mathbf{y}_i^{\text{eq}}$ its scale-steered RGB-logit readout. Sampling, padding, clipping, aliasing, and cross-slot competition are the finite-grid approximation boundaries.

The Obj3D instantiation broadcasts each slot to 16×16. It applies bilinear $2\times$ synthesis followed by a 5×5 scale-steered convolution at 32×32, another bilinear synthesis followed by a 3×3 scale-steered convolution at 64×64, and three 3×3 scale-steered blocks at the final resolution. The alpha head is 1×1. The RGB path pointwise-fuses the equivariant field with a spatially constant appearance style before its scale-steered head; only its bounded, zero-initialized micro branch uses a fixed-pixel 3×3 convolution. Progressive synthesis avoids the checkerboard phase preference of zero-insertion while retaining learned scale-dependent aggregation after every resolution change.

The global reference scale is continued during cold start without breaking the homogeneity of Eq. (6). Abbreviate $s_c = s_{\text{cold}}$ and $s_r = s_{\text{ref}}$. With update t,

$$s_{\text{ref}}^{\text{eff}}(t) = \begin{cases} s_c, & 0 \le t < 2{,}000, \\ s_c^{1-\beta_t} s_r^{\beta_t}, & 2{,}000 \le t < 10{,}000, \\ s_r, & t \ge 10{,}000, \end{cases} \tag{C.12}$$

where $\beta_t = \text{clip}_{[0,1]}((t - 1999)/8000)$, s_{cold} is the RMS scale of uniform attention on the initial grid, and s_{ref} is the final operator gauge. At every fixed update t, $\rho_{\ell,t}(ks) = k\rho_{\ell,t}(s)$ still holds for every layer ℓ.

C.6 Decoder Implementations

We evaluate scale-steered and canonical-canvas slot decoding within the GeoCo-SAVi geometry-control formulation. The canonical-canvas implementation follows the local-patch transformation approach of AIR and SPACE (Eslami et al., 2016; Lin et al., 2020b). The decoder produces RGB and alpha logits on a fixed 64×64 canvas at $p = 0$ and $s_0 = 0.2$, using ordinary convolutions of the same widths and kernel sizes. A final bilinear warp samples at $(\mathbf{x} - p)s_0/s$, with aligned corners; out-of-canvas RGB and alpha logits are filled with 0 and -20, respectively. RGB sigmoid and slotwise alpha normalization follow the warp. Both use factual and counterfactual calibration.

Table 5: Obj3D decoder implementations within GeoCo-SAVi at 70k updates. $\mathrm{PSNR}_{\mathrm{pos/scl/app}}$ measures full-image agreement with independent RGB edit targets (Sec. E.6). F-scores are percentages; PSNR is in dB.

Decoder	PSNR↑	$\mathrm{PSNR}_{\mathrm{pos}}\uparrow$	$\mathrm{PSNR}_{\mathrm{scl}}\uparrow$	$\mathrm{PSNR}_{\mathrm{app}}\uparrow$	$F_{\mathrm{pos}}\uparrow$	$F_{\mathrm{scl}}\uparrow$	$F_{\mathrm{app}}\uparrow$
Scale-steered	44.210	39.914	40.176	35.728	93.29	92.72	91.09
Canonical-canvas	43.501	39.568	39.945	36.148	93.00	93.94	89.17

Decoder	$E_{p\to c^\star}\downarrow$	$E_{\hat{c}\to c^\star}\downarrow$	$O_{\mathrm{attn}}\downarrow$	$E_{\Delta p}\downarrow$	$\beta_r \to 1$	$\beta_A \to 2$	$D_{\mathrm{app}}\downarrow$
Scale-steered	0.0436	0.0391	0.0007	0.0051	1.0339	2.0225	0.0124
Canonical-canvas	0.0451	0.0428	0.0007	0.0046	0.9763	1.9315	0.0097

The canonical-canvas implementation achieves higher transplant oracle PSNR and F_{scl}, with lower translation error and transplant drift. The scale-steered implementation achieves higher reconstruction PSNR, position and scale oracle PSNR, F_{pos}, and F_{app}. We use the scale-steered implementation in the main experiments to balance reconstruction quality and geometric control.

C.7 Appearance-Transplant Pairing

Each recipient scene replaces only one foreground slot, avoiding simultaneous changes to several footprints and their background competition. Pair eligibility is computed entirely from detached factual decompositions. We exclude inactive or degenerate supports, strong boundary truncation, and excessive foreground overlap. For MOVi-C, a detached factual-coherence refinement additionally requires OneFG compactness at least 1, at least 8 effective attention pixels, and IoU at least 0.6 between hard slot ownership and its gamma-2 OneFG support thresholded at 0.5. Obj3D uses no additional coherence refinement. The pair confidence ω_{ij} combines the base factual quality terms; it never observes the counterfactual outcome, so a poor transplant cannot escape supervision by invalidating its own pair.

This self-supervised factual filtering constructs training losses, where privileged input-side object cores are unavailable. Evaluation populations and metric denominators are specified separately in Sec. E.

The compactness target in the main paper implies the coverage relation

$$A_{i\leftarrow j}^{\mathrm{target}} = \mathrm{sg}\big(K(m_j^{\mathrm{don}})\big)\,\mathrm{sg}(r(m_i^{\mathrm{rec}}))^2\,. \tag{C.13}$$

Thus recipient geometry commands spatial extent, while donor appearance may retain its fill density, transparency, and occlusion behavior.

C.8 Single-Frame Schedule

The Obj3D curriculum uses 70k steps of single-frame training with random position initialization. Optimization uses AdamW with batch size 64, base learning rate 4×10^{-4}, zero weight decay, BF16 AMP, global gradient-norm clipping at 0.05, a 10k-step linear warm-up followed by cosine decay to zero at 70k, and decoder compilation:

- steps 0–15k: reconstruction warm-up, with the terminal-gate regularizer;
- steps 15k–40k: ramp and retain $\mathcal{L}_{\mathrm{pos}}$ and $\mathcal{L}_{\mathrm{ov}}$;
- steps 40k–50k: retain factual losses and add counterfactual appearance transplantation with $\mathcal{L}_{\mathrm{geo}}$, using the coordinate-wise center penalty;

- steps 50k–60k: retain these objectives and switch to the scale-normalized center penalty;
- steps 60k–70k: additionally enable $\mathcal{L}_{\text{tail}}$ with $\lambda_{\text{tail}} = 10^{-3}$, without an additional ramp.

The center and tail penalties are defined in Sec. C.3. The reported Obj3D configuration uses $\lambda_{\text{pos}} = 0.2$, $\lambda_{\text{ov}} = 0.01$, $\lambda_{\text{geo}} = 0.002$, $\lambda_K = 0.25$, and $\lambda_{\text{gate}} = 10^{-4}$. Auxiliary losses other than the late tail penalty enter through ramps near their activation boundaries. In the main trajectory, the optional selector is jointly trained from 35k with its disjoint AdamW optimizer (Loshchilov & Hutter, 2019) and is applied from 36.2k with a ramped transfer cap. Its threshold and cap were selected on a calibration panel drawn from the training split; no validation or evaluation sample entered this panel. The selected operating point is then used for subsequent training. The 40k diagnostic in Fig. 4 comes from a separate controlled run using reconstruction, $\mathcal{L}_{\text{pos}}$, and $\mathcal{L}_{\text{ov}}$ only. The diagnostic omits teacher–selector activation, counterfactual geometry, and temporal initialization. The selector's class/ranking weights are $\lambda_{-} = 2$ and $\lambda_{\text{rank}} = 0.25$.

C.9 Temporal Initializer Semantics

The explicit update is

$$\begin{aligned}(\delta\mathbf{a}, \delta\mathbf{p}, \delta\ell)_{t,i} &= \mathbf{W}_{\text{init}}\big(\mathcal{F}_\Theta(\mathbf{Z}_{<t})\big)_i,\\ \mathbf{a}^{(0)}_{t,i} &= \mathbf{a}_{t-1,i} + \delta\mathbf{a}_{t,i},\\ \mathbf{p}^{(0)}_{t,i} &= \text{clip}(\mathbf{p}_{t-1,i} + \delta\mathbf{p}_{t,i}),\\ s^{(0)}_{t,i} &= \text{clip}(s_{t-1,i}\exp(\delta\ell_{t,i})).\end{aligned} \tag{C.14}$$

The temporal module consumes the raw slot state $\mathbf{z} = [\mathbf{a}; \mathbf{p}; s]$ and predicts an initialization residual, not a separately evaluated future state. Its zero-initialized final projection returns $(\delta\mathbf{a}, \delta\mathbf{p}, \delta\ell)$: appearance uses an additive update, position a bounded additive update, and scale $s^{(0)} = \text{clip}(s\exp(\delta\ell))$. Thus the update is additive in log-scale even though the Transformer input is raw scale. The residual adjusts appearance and geometry for next-frame ISA refinement. The time–slot Transformer (Li et al., 2025; Vaswani et al., 2017) preserves the slot axis, mixes scene context through spatial attention, and returns one residual per slot. Ordered carry-forward provides slot-wise temporal correspondence for current-frame ISA refinement. Next-frame reconstruction jointly trains the Transformer and final projection through current-frame ISA refinement. Freezing or jointly adapting the calibrated single-frame core is a benchmark protocol choice.

C.10 Gradient Routing Summary

Gradient routing is defined as follows:

- $\mathcal{L}_{\text{rec}}$ updates the full core.
- $\mathcal{L}_{\text{ov}}$ keeps attention live.
- $\mathcal{L}_{\text{pos}}$ keeps the OneFG target and selected-background logit paths live but detaches the position and scale commands, background-slot identities, and factual validity mask.
- $\mathcal{L}_{\text{geo}}$ keeps the counterfactual alpha path, donor appearance, and recipient factual background logits live while detaching recipient commands, factual moments, background-slot identities, pair selection, and confidence.
- $\mathcal{L}_{\text{tail}}$ detaches its threshold mask, foreground/background identities, and pair-confidence weights, while keeping target and background logits live. Its factual term follows the detached-command position-loss route; its counterfactual term uses donor appearance with detached recipient commands and the recipient's live factual background logits.
- teacher construction, selector inputs, and deployment eligibility are detached; $\mathcal{L}_{\text{sel}}$ updates only selector parameters through a disjoint optimizer.

D Shared Evaluation Metrics and Notation

This section collects the metric notation shared by Obj3D and MOVi-C. The populations, support roles, boundary scopes, and aggregation rules are stated in the corresponding evaluation sections.

Decoded-support geometry. For a hard decoded support $\widehat{M}_i$ on the endpoint-aligned grid $\mathbf{q}(\mathbf{u}) \in [-1,1]^2$, let

$$n_i = \sum_{\mathbf{u}} \widehat{M}_i(\mathbf{u}), \qquad \widehat{\mathbf{c}}_i = \frac{\sum_{\mathbf{u}} \widehat{M}_i(\mathbf{u})\mathbf{q}(\mathbf{u})}{n_i},$$
$$\widehat{r}_i^2 = \frac{\sum_{\mathbf{u}} \widehat{M}_i(\mathbf{u})\|\mathbf{q}(\mathbf{u}) - \widehat{\mathbf{c}}_i\|_2^2}{n_i}, \qquad \widehat{A}_i = \frac{n_i}{HW}. \tag{D.1}$$

Thus $\widehat{\mathbf{c}}_i$, $\widehat{r}_i$, and $\widehat{A}_i$ are the decoded centroid, normalized-grid RMS radius, and coverage fraction. Centroids are defined for supports with positive mass; radius and coverage are zero for zero-mass supports.

Factual geometry and allocation. Let $\mathbf{p}_i$ be the explicit position and $\mathbf{c}_i^\star$ an external reference center. The three factual position errors are

$$E_{p\to\hat{c}} = \|\mathbf{p}_i - \widehat{\mathbf{c}}_i\|_2, \quad E_{p\to c^\star} = \|\mathbf{p}_i - \mathbf{c}_i^\star\|_2,$$
$$E_{\hat{c}\to c^\star} = \|\widehat{\mathbf{c}}_i - \mathbf{c}_i^\star\|_2. \tag{D.2}$$

All three use L2 distance on the normalized $[-1,1]^2$ grid. For the set $\mathcal{I}_x$ of attention maps participating in an example or frame x, write $N_x = |\mathcal{I}_x|$ and spatially normalize each map so that $\sum_n w_i(\mathbf{u}_n) = 1$. Normalized attention overlap is

$$O_{\mathrm{attn}}(x) = \frac{1}{N_x(N_x-1)} \sum_{\substack{i,j\in\mathcal{I}_x \\ i\neq j}} \sum_n w_i(\mathbf{u}_n) w_j(\mathbf{u}_n). \tag{D.3}$$

The reported O_{attn} averages this ordered off-diagonal overlap over the stated evaluation population.

At fixed geometry $g = (\mathbf{p}, s)$, pooling all eligible decoded appearances gives, for $x \in \{\widehat{r}, \widehat{A}\}$,

$$\sigma_{x|g} = \sqrt{J^{-1}\sum_{j=1}^{J}(x_j - \bar{x})^2}. \tag{D.4}$$

Here J is the number of defined measurements in the full population. Each object–frame record has equal weight; $\bar{x}$ is its population mean.

Editing command adherence. For an edited mask M_e, let M_t be the corresponding factual decoded mask transformed by the editing command. Editing F1 is

$$F_1(M_e, M_t) = \frac{2|M_e \cap M_t|}{|M_e| + |M_t|}. \tag{D.5}$$

F_{pos}, F_{scl}, and F_{app} denote the corresponding averages for position, scale, and appearance transplantation. F1 is evaluated on valid edits (Sec. E); an empty edited support receives zero. For a commanded pixel translation $\mathbf{d}_{\mathrm{px}}$, translation error is

$$E_{\Delta p} = \frac{\|(\mathbf{c}_{\mathrm{px}}^e - \mathbf{c}_{\mathrm{px}}^f) - \mathbf{d}_{\mathrm{px}}\|_2}{\sqrt{H^2 + W^2}}, \tag{D.6}$$

where superscripts f and e denote factual and edited outputs. Appearance-transplant drift is

$$D_{\mathrm{app}} = \|\widehat{\mathbf{c}}_r^e - \widehat{\mathbf{c}}_r^f\|_2, \tag{D.7}$$

measured on the normalized grid while recipient geometry is fixed.

For a five-command scale curve, ordinary least squares with an intercept fits $\log \widehat{r}(k)$ or $\log \widehat{A}(k)$ against $\log k$. The resulting slopes β_r and β_A have ideal values one and two.

For a video, the tube form pools intersection and support area over its T frames before taking the ratio:

$$F_{\mathrm{tube}} = \frac{2\sum_{t=1}^{T}|M_t^e \cap M_t^{\mathrm{ideal}}|}{\sum_{t=1}^{T}(|M_t^e| + |M_t^{\mathrm{ideal}}|)}. \tag{D.8}$$

E OBJ3D EVALUATION PROTOCOL AND EXTENDED RESULTS

E.1 MODELS, VALIDATION POPULATION, AND SUPPORT ROLES

The ISA baseline uses the scalar isotropic scale in Eq. (5). ISA, SlotAug, and GeoCo-SAVi use six slots with 64-dimensional appearance states and matched encoder/decoder capacity, training exposure, and RGB reconstruction resolution. SlotAug retains its published objectives and commands; command normalization is applied in a subsequent hidden layer to preserve instruction information. Reconstruction is evaluated with four inference initializations. The geometry experiments below use the first frame of all 2,600 validation windows, batch size four, and initialization seed $42 + b$ for batch index b. The validation-loader seed is 20260717.

Training support is obtained from the model's decoder and attention outputs. *Measurement support* is the decoded mask,

$$\widehat{M}_i(\mathbf{u}) = \mathbb{I}[i = \arg\max_j \alpha_j(\mathbf{u})].$$

Reference support denotes an RGB pseudo-label or a visible GT mask used for the separate external position evaluation. Measured centroid, radius, and coverage follow Eq. D.1 on the endpoint-aligned grid. The background slot maximizes factual mean alpha. We consider objects with at least ten native-canvas support pixels to be valid, excluding negligible supports. Valid edits additionally require the ideal transformed support to meet the same criterion. Evaluation retains valid non-background objects on the native 64×64 canvas. Background identity and eligibility are fixed before intervention.

We report two scopes: *all valid objects*, and *ideal-boundary-feasible objects*. The latter includes valid edits whose factual and ideal transformed supports both avoid the outermost pixel band. Feasibility is evaluated on transformed occupied pixel-cell bounds before canvas clipping. For transplantation, both donor and recipient factual supports and the transformed donor must satisfy this condition. A scale curve is boundary-feasible only when all five commands are feasible. Actual edited masks do not determine eligibility.

E.2 EDITING COMMANDS AND AGGREGATION

All interventions decode frozen factual slots. Translation uses four pixel commands $(\pm 6, 0)$ and $(0, \pm 6)$, converted to grid displacements $\boldsymbol{\delta} = 2\mathbf{d}_{\mathrm{px}}/63$. Scaling uses $k \in \{0.5, 0.75, 1, 1.25, 1.5\}$. The ideal coordinate maps are

$$T_{\mathrm{pos}}(\mathbf{q}) = \mathbf{q} + \boldsymbol{\delta}, \qquad T_{\mathrm{scl}}(\mathbf{q}) = \mathbf{p}_i + k(\mathbf{q} - \mathbf{p}_i).$$

ISA and GeoCo-SAVi modify the explicit position or scale; SlotAug uses its native command-and-identity decoding path. Its scale translation component compensates for the difference between its attention centroid and the explicit position, so the ideal scale target is centered at $\mathbf{p}_i$.

For appearance transplantation, the donor is the next window from a different source in cyclic validation order, with the same foreground-slot rank. Both donor and recipient must be valid objects. The donor's 64 appearance coordinates replace the recipient's appearance while recipient $(\mathbf{p}_r, s_r)$ and other slots are fixed. This is a representation-level transplantation diagnostic for SlotAug as well. The ideal donor-mask map is

$$T_{\mathrm{app}}(\mathbf{q}) = \mathbf{p}_r + \frac{s_r}{s_d}(\mathbf{q} - \mathbf{p}_d).$$

Ideal binary masks are inverse-warped with nearest-neighbor sampling, endpoint alignment, and zero padding. Donor cropping and occlusion remain in its factual support.

Editing F-scores follow Eq. D.5. Scale F1 averages the four nonidentity commands. The no-op baseline scores the unchanged recipient mask against the target of the requested edit. Main-text scores average valid edits without boundary filtering; they measure adherence to the prescribed editing commands.

Position errors, appearance drift, and translation error follow Eqs. D.2–D.6. On a square endpoint-aligned canvas, a diagonal-normalized distance converts to grid L2 by multiplication by $2\sqrt{H^2 + W^2}/(W - 1)$. PSNR uses dB, radius uses normalized-grid RMS-radius units, and coverage is an area fraction.

Scale-response slopes follow Sec. D. Slopes require positive values at every command. Center errors use outputs with defined centroids; their counts are reported alongside F1 totals. Radius and coverage retain zero values for zero-area outputs.

E.3 Fixed-Geometry Stability

Each valid non-background factual slot is decoded with its factual background slot and attention held fixed. The target is set to $\mathbf{p} = 0$ and $s \in \{0.1, 0.2, 0.3\}$. The pooled radius and coverage standard deviations follow Eq. D.4. Radius and coverage include zero-area outputs with value zero. No edited-support size or boundary test is applied. Attention overlap is measured separately over all factual normalized attention maps using Eq. D.3.

Shared-background evaluation. We also evaluate cross-appearance stability with a shared factual background. For each method, we sample 1,024 valid factual objects and eight background slots from distinct videos using a fixed random seed; the background source videos are shared across methods within each dataset. Background slots follow the same factual mean-alpha rule above. Each target is decoded with one background slot, whose representation and attention are held fixed across the sampled appearances. Target attention is retained, with $\mathbf{p} = 0$ and $s \in \{0.1, 0.2, 0.3\}$. All outputs contribute to the statistics, with zero radius and coverage for zero-area supports.

For each background, we compute the mean, standard deviation, and coefficient of variation (CV) across appearances, then average these statistics over backgrounds. Table 6 reports $s = 0.2$. GeoCo-SAVi has lower average within-background standard deviations and CVs for radius and coverage than the corresponding baselines at all three tested scales.

Table 6: Fixed-geometry stability with shared factual backgrounds at $s = 0.2$. Each statistic is computed across 1,024 appearances within a background and averaged over eight backgrounds.

		Radius			Coverage		
Dataset	Method	Mean	Std.↓	CV↓	Mean	Std.↓	CV↓
Obj3D	ISA	0.15430	0.04294	0.27831	0.03656	0.01793	0.49034
	SlotAug	0.14300	0.04978	0.34815	0.03266	0.02145	0.65703
	GeoCo-SAVi	0.15521	**0.02029**	**0.13075**	0.03364	**0.00836**	**0.24846**
MOVi-C	STAITUS	0.25065	0.09912	0.49495	0.10652	0.07563	0.88901
	GeoCo-SAVi	0.14799	**0.04616**	**0.32078**	0.02830	**0.01491**	**0.54265**

E.4 External RGB Reference Evaluation

Obj3D supplies no official instance masks. A model-independent RGB bank provides instance masks. Color and temporal cues identify object regions; RGB-based refinement separates touching instances and joins fragmented supports. A one-pixel boundary allowance covers object-edge pixels.

The bank provides 9,265 objects across all 2,600 validation windows. Hungarian maximum-overlap association maps the reference objects to factual decoded slots. The associated slot's position and full decoded centroid are compared with the reference-mask centroid. The shared evaluation set contains 9,255 objects with defined centroids across the compared methods and controls. Table 7 gives the three distances on this population; the all-slot position measurement is reported separately.

E.5 Extended Numerical Results

Tables 9–14 report editing, scale response, sample counts, fixed-geometry variation, and no-op scores. Both scopes retain all eligible actual outcomes. F1 additionally reports valid edits in the interior subset. Section G reports factual scale calibration on both datasets.

F-score confidence intervals use 10,000 source-video bootstrap replicates and the 2.5th and 97.5th percentiles, preserving each reported score's aggregation. All observations from a video share its resampling weight; transplant pairs receive the product of recipient and donor video weights from the same resample. The intervals summarize evaluation-sample variability.

Table 7: External position evaluation in normalized-grid L2 units: Obj3D uses 9,255 RGB pseudo-reference objects from all 2,600 windows, and MOVi-C uses 1,543 visible GT tracks. Within-model distance is evaluated on the same associated objects.

Dataset / method	$E_{p\to\hat{c}}$	$E_{p\to c^\star}$	$E_{\hat{c}\to c^\star}$
Obj3D / ISA	0.10135	0.10232	0.04204
Obj3D / SlotAug	0.04658	0.05794	0.03584
Obj3D / GeoCo-SAVi	0.01126	0.04361	0.03910
MOVi-C / STAITUS	0.08683	0.34170	0.31277
MOVi-C / GeoCo-SAVi	0.02761	0.24032	0.23406

Table 8: All-slot position error in normalized-grid L2 units. Position uses original factual ownership; video editing uses a consistent decoder precision for factual reference and intervention, so valid-object counts can differ slightly.

Dataset / method	$E_{p\to\hat{c}}$	Defined	Edit objects
Obj3D / ISA	0.10489	9424	9424
Obj3D / SlotAug	0.04763	9357	9357
Obj3D / GeoCo-SAVi	0.01065	9320	9320
MOVi-C / STAITUS	0.06865	51962	51971
MOVi-C / GeoCo-SAVi	0.02567	51600	51629

Table 9: Obj3D editing results. F1 (%) uses valid edits in each scope; interior uses factual/ideal feasibility. Brackets give 95% confidence intervals for the main-table F-scores. $E_{\Delta p}$ is a diagonal fraction and D_{app} is grid L2.

Method / scope	F_{pos}	F_{scl}	F_{app}	$E_{\Delta p}$	D_{app}
ISA (all)	86.35	62.84	75.73	0.01012	0.05410
	[85.84, 86.83]	[62.40, 63.28]	[74.44, 76.93]		
SlotAug (all)	90.78	82.19	78.20	0.00770	0.06012
	[90.34, 91.22]	[82.03, 82.34]	[76.79, 79.49]		
GeoCo-SAVi (all)	93.29	92.72	91.09	0.00645	0.02083
	[93.03, 93.55]	[92.60, 92.83]	[90.43, 91.64]		
ISA (interior)	85.83	61.08	78.79	0.00908	0.04422
SlotAug (interior)	91.19	81.75	84.65	0.00599	0.03214
GeoCo-SAVi (interior)	93.40	93.20	93.08	0.00509	0.01240

Table 10: Obj3D five-point scale response. Slopes are means over defined curves; counts are defined/total.

Method / scope	β_r	β_A	β_r def.	β_A def.
ISA (all)	0.09882	-0.01512	9344/9424	9370/9424
SlotAug (all)	0.54612	1.04735	9339/9357	9344/9357
GeoCo-SAVi (all)	1.03473	1.98611	9303/9320	9310/9320
ISA (interior)	0.09388	-0.00923	6928/6952	6939/6952
SlotAug (interior)	0.54118	1.05748	7310/7314	7311/7314
GeoCo-SAVi (interior)	1.03391	2.02246	7380/7386	7384/7386

E.6 Object-Matched Editing and RGB Preservation

We use the RGB instance bank above to compare edits of the same reference objects across ISA, SlotAug, and GeoCo-SAVi. Factual maximum-overlap association fixes the slot for each reference

Table 11: Obj3D scope-specific F1 denominators and defined/total center-error counts. F1 retains all actual outputs for valid edits, including empty outputs.

Method / scope	N_{pos}	N_{scl}	N_{app}	$E_{\Delta p}$ def.	D_{app} def.
ISA (all)	37144	36631	7402	37511/37696	7417/7418
SlotAug (all)	36843	36619	7691	37068/37428	7689/7691
GeoCo-SAVi (all)	36616	36858	7397	37221/37280	7397/7397
ISA (interior)	30842	30357	4564	30831/30842	4564/4564
SlotAug (interior)	31390	31258	5293	31328/31390	5291/5293
GeoCo-SAVi (interior)	30319	30654	4698	30319/30319	4698/4698

Table 12: Obj3D conditional output diagnostics, shown as count/eligible count (%). Disappearance counts empty outputs for positive-area ideal targets.

Method / scope	Position	Scale	Transplant
ISA (within-canvas)	11/30842 (0.04)	69/30357 (0.23)	0/4564 (0.00)
SlotAug (within-canvas)	62/31390 (0.20)	8/31258 (0.03)	2/5293 (0.04)
GeoCo-SAVi (within-canvas)	0/30319 (0.00)	4/30654 (0.01)	0/4698 (0.00)
ISA (all)	79/37358 (0.21)	70/37696 (0.19)	1/7418 (0.01)
SlotAug (all)	78/37064 (0.21)	16/37428 (0.04)	2/7691 (0.03)
GeoCo-SAVi (all)	1/36805 (0.00)	12/37280 (0.03)	0/7397 (0.00)

Table 13: Obj3D pooled cross-appearance fixed-geometry statistics. The radius coefficient of variation is $\mathrm{CV}_{r|g} = \sigma_{r|g}/\bar{r}$. At $s = 0.2$, GeoCo-SAVi and ISA have similar mean radii, with lower radius standard deviation and CV for GeoCo-SAVi. N counts all measurements.

Method	s	$\bar{r}$	$\sigma_{r\|g}$	$\mathrm{CV}_{r\|g}$	$\sigma_{A\|g}$	N
ISA	0.1	0.14826	0.03351	0.22600	0.01239	9424
ISA	0.2	0.15308	0.04405	0.28775	0.01833	9424
ISA	0.3	0.15779	0.07107	0.45039	0.02911	9424
SlotAug	0.1	0.12094	0.02476	0.20475	0.00874	9357
SlotAug	0.2	0.14279	0.05201	0.36426	0.02265	9357
SlotAug	0.3	0.10030	0.10388	1.03572	0.04049	9357
GeoCo-SAVi	0.1	0.07268	0.01056	0.14530	0.00208	9320
GeoCo-SAVi	0.2	0.15386	0.02109	0.13709	0.00856	9320
GeoCo-SAVi	0.3	0.23137	0.03109	0.13439	0.01907	9320

Table 14: Obj3D no-op F1 (%): unchanged recipient masks scored against the requested edit targets, on the same valid edits without boundary filtering.

Method	$F^{\text{noop}}_{\text{pos}}$	$F^{\text{noop}}_{\text{scl}}$	$F^{\text{noop}}_{\text{app}}$
ISA	34.62	62.54	69.51
SlotAug	33.56	63.54	77.83
GeoCo-SAVi	37.02	63.54	80.04

object. The commands are the four translations and four nonidentity scales in Sec. E.2. All results in this subsection average edits within each source video and then equally across the 200 videos.

Common-object editing. The common pool requires valid reference edits and valid edits for all three associated decoded supports. It is fixed from factual supports and the prescribed transformations. Editing F1 follows Eq. D.5 and the coordinate maps in Sec. E.2. Table 15 reports 36,210 translation and 35,986 scale edits, each spanning 9,216 reference objects and all 2,600 validation windows.

Table 15: Obj3D command adherence on common reference object–command pairs. F1 is in percent; each column uses the same pairs for all methods.

Method	$F_{\mathrm{pos}} \uparrow$	$F_{\mathrm{scl}} \uparrow$
ISA	88.65	63.85
SlotAug	91.98	82.62
GeoCo-SAVi	**93.77**	**92.93**

Independent RGB targets. An RGB visible-layer oracle provides a common target image for each reference object and command. We extract the object using its RGB-derived matte, fill its source region by local background inpainting, and composite the transformed premultiplied RGB layer with bilinear sampling. Scaling is centered at the reference-mask centroid. These target images are constructed from the input RGB and reference masks independently of the model predictions. We evaluate all eight commands for each of 9,240 valid reference objects across the 2,600 windows, giving 73,920 edits per method. Table 16 reports PSNR on the full 64×64 image and on the union of the source and transformed reference supports, expanded by two four-connected pixel steps.

For transplantation, each valid reference object is paired with a donor from the next video in cyclic order, giving 9,240 common pairs across 200 recipient videos. The visible donor layer is aligned to the recipient reference-mask centroid and RMS radius, then composited over the inpainted recipient image. The associated donor appearance state replaces the recipient appearance state, with geometry and other slots fixed. All pairs are retained without boundary filtering. The editing region includes the recipient and transformed donor supports and alpha footprints, expanded as above. Scores are averaged within recipient videos and then equally over videos.

Table 16: Obj3D PSNR against independent RGB oracle edits (dB). Position and scale each contain 36,960 edits; transplant contains 9,240 pairs. Target images are identical across methods.

	Full image			Editing region		
Method	Position↑	Scale↑	Transplant↑	Position↑	Scale↑	Transplant↑
ISA	37.322	35.368	34.606	27.675	24.349	23.327
SlotAug	38.654	38.528	34.951	29.720	28.924	24.047
GeoCo-SAVi	**39.914**	**40.176**	**35.728**	**30.883**	**30.889**	**24.428**

Non-target RGB preservation. On the position and scale object–command pairs, we measure RGB changes outside the editing region defined above. This region is shared across methods. For RGB intensities in $[0, 1]$, we report the mean absolute change from the factual reconstruction and the percentage of pixels whose channel-averaged absolute change exceeds $2/255$. GeoCo-SAVi has lower values for both measures (Table 17). Paired 95% confidence intervals from 5,000 source-video bootstrap replicates exclude zero for the reported improvements over both baselines in Tables 15–17.

Table 17: Obj3D RGB changes outside the source/target editing region. MAE is multiplied by 10^3; changed-pixel rates are percentages.

	RGB MAE $\times 10^3$		Changed pixels (%)	
Method	Position↓	Scale↓	Position↓	Scale↓
ISA	0.8902	0.7724	2.1480	2.0424
SlotAug	0.8296	0.7756	2.1646	1.9778
GeoCo-SAVi	**0.4939**	**0.6873**	**1.2360**	**1.6966**

E.7 Component Comparisons

The seven controls in Table 4 are evaluated at 70k updates. The conventional decoder replaces scale-steered spatial convolutions with ordinary unit-grid convolutions, retaining the relative-coordinate input, feature widths, upsampling stages, and output heads.

The loss controls set the corresponding objective weight to zero: $\mathcal{L}_{\mathrm{ov}}$, $\mathcal{L}_{\mathrm{pos}}$, $\mathcal{L}_{\mathrm{geo}}$, or $\mathcal{L}_{\mathrm{tail}}$. Removing $\mathcal{L}_{\mathrm{geo}}$ disables its center, radius, and compactness terms; the counterfactual branch remains available for the tail penalty. The readout control disables the terminal readout and its gate regularizer. The selector control disables the alpha selector. For active losses, factual calibration and overlap regularization begin at 15k, and counterfactual calibration begins at 40k with the ramps specified above. The counterfactual center penalty uses coordinate Huber before 50k and the scale-normalized form thereafter. The tail penalty, when active, has weight 0.001 and threshold 0.25 from 60k without a ramp. These settings complement the component-specific comparisons in the main text.

PSNR averages four inference initializations. Editing uses the same 2,600 validation windows and valid-object criterion as the main evaluation. Reference-center errors use the fixed RGB reference bank described above; fixed-geometry spreads pool objects at $p = 0, s = 0.2$.

F MOVi-C Protocol and Extended Diagnostics

F.1 Common Video Protocol and Reference Implementations

All three methods are trained on MOVi-C (Greff et al., 2022) for 100k optimizer updates and evaluated on the fixed official validation population of 250 videos, each containing 24 frames. RGB reconstruction uses 64×64 frames; instance metrics use the official GT resolution after alpha resizing, and editing protocols state their measurement resolution below. Native RGB is first resized bilinearly to 64×64, then bicubically to 336×336 for the frozen visual encoder `vit_small_patch14_dinov2.lvd142m` (Oquab et al., 2024); its block-12 384-dimensional tokens are projected to 128 dimensions. The reference implementations are matched in data, RGB target, encoder/decoder capacity, batch size, 100k update budget, and total sample exposure. Loss, optimizer, and curriculum hyperparameters otherwise use the published values, subject only to the adaptations disclosed below.

F.2 GeoCo-SAVi MOVi-C Architecture, Optimization, and Curriculum

Slot core and renderer. GeoCo-SAVi uses 11 slots with a 64-dimensional appearance state and explicit state $(\mathbf{a}, \mathbf{p}, s)$. It uses 64-dimensional query/key/value projections, a 64-dimensional Slot-Attention GRU (Locatello et al., 2020; Cho et al., 2014), a 256-dimensional update MLP, and a 128-dimensional grid encoding. The first frame of every training clip uses three Slot-Attention iterations; a carry-forward frame uses two. Positions are initialized iid-uniformly, s is clamped to $[.001, 2]$, factor 5 normalizes the object-relative coordinate field, and the Spatially Equivariant Decoder has width 64. It forms a 32×32 canonical field, applies one scale-steered 5×5 refinement, bilinearly upsamples, then applies four scale-steered 3×3 refinements at 64×64. The RGB head is 3×3 with sigmoid output and the alpha head is 1×1. The reported renderer uses only the convolutional path described above. The regular steered kernel is 3×3 and its learned gain is bounded by 2.

Optimization and reference-scale continuation. We use AdamW with base learning rate 4×10^{-4}, zero weight decay, BF16 AMP, global gradient-norm clipping at 0.05, physical/effective batch size 16. Only the decoder is compiled. The core rate warms up linearly for 10k updates and then follows cosine decay to zero at 100k. The training loader has eight workers, no subsampling, and uses video identity as its source key; the fixed position seed for evaluation is 20260723. Let $s_c = 0.851256531$, $s_f = 0.2$, and $\ell_t = (1 - \beta_t) \log s_c + \beta_t \log s_f$. The global numerical reference of the scale-steered operator follows

$$s_{\mathrm{ref}}(t) = \begin{cases} s_c, & t < 2000, \\ e^{\ell_t}, & 2000 \le t < 10000, \\ s_f, & t \ge 10000, \end{cases} \tag{F.1}$$

where $\beta_t = \min((t - 1999)/8000, 1)$. This reference is a global numerical continuation for the convolution operator, not an object-size prior or a replacement for the predicted s_i.

Loss activation and temporal curriculum. Position and normalized-attention overlap activate at 15k and reach full weight over 2k updates. We report the rounded coefficients $\lambda_{\text{pos}} = 0.04$ and $\lambda_{\text{ov}} = 0.018$. Position uses coordinate Huber loss with threshold 0.05 on the detached-command gamma-2 OneFG centroid; only loose factual foreground supports with OneFG coverage at most 0.2 pass the detached validity filter, and no separate valid-background gate is applied. $\mathcal{L}_{\text{geo}}$ is kept at zero until 48k, ramps linearly to full weight by 50k, and then uses $\lambda_{\text{geo}} = 0.004$, unit center/radius terms, and a 0.25 compactness term, all with Huber threshold 0.05. Its gamma-2 OneFG pairs use factual quality gates and the receiver-center/radius, donor-compactness semantics described in Sec. 3. It is evaluated at time index 0 for two-frame clips and at indices 0 and 2 for four-frame clips.

Table 18: Reported GeoCo-SAVi MOVi-C curriculum. All stages share one optimizer, scheduler, and AMP-scaler trajectory.

Updates	Frames / clip	Active terms	Purpose
0–15k	2	RGB reconstruction; s_{ref} continuation	Stable slot formation with carry-forward
15–17k	2	$\mathcal{L}_{\text{pos}}$, $\mathcal{L}_{\text{ov}}$ ramp	Factual geometry and allocation calibration
17–48k	2	$\mathcal{L}_{\text{pos}}$, $\mathcal{L}_{\text{ov}}$	Core-only stabilization
48–50k	2	$\mathcal{L}_{\text{geo}}$ ramp	Counterfactual geometry activation
50–70k	2	Reconstruction, factual and counterfactual objectives	Appearance-transplantation calibration
70–72k	4	Zero-residual STATM local LR warm-up	Temporal-module activation
72–100k	4	Full STATM local schedule	Joint temporal refinement

The residual STATM initializer has width 128, four heads, a 512-dimensional MLP, and four history states. It detaches history and uses zero dropout. Its output projection is zero-initialized, so the transition at 70k starts from exact carry-forward. Its parameter group belongs to the same AdamW instance from update zero but receives no gradient before 70k; it then warms up locally through 72k to 0.25 times the base rate, then follows an independent cosine decay to zero at 100k. The reported MOVi-C configuration combines the spatially equivariant core with the temporal initializer; the optional terminal appearance readout, selective alpha ownership, and tail penalty are inactive.

The STAITUS implementation uses the same-capacity SAVi encoder/decoder core, a temporal appearance predictor, and an active-set gate, with the published losses. The gate is a part of reconstruction and the losses, with a 15k all-active warm-up followed by a 15k gate ramp. STATM-SAVi uses the same-capacity SAVi encoder/decoder core, RGB MSE, and the published corresponding-slot spatial and temporal Transformer in its strongest CS/T+S form. These are same-protocol implementations for controlled comparison.

F.3 Factual Reconstruction and Instance Metrics

Full-softmax alpha is bilinearly resized to the official GT resolution before taking a slotwise argmax. FG-ARI, global ARI, Hungarian mIoU, and mean best overlap (mBO) are averaged with equal weight per video; global ARI retains the GT background. RGB values lie in $[0, 1]$; PSNR is $10 \log_{10}(1/\text{MSE})$ per 24-frame video, with MSE averaged over all RGB entries in that video and the 250 video-level scores weighted equally. Neither the factual metrics nor their denominators use quality filters, OneFG, explicit geometry, or edited states. Official visible instance segmentations are used only for these metrics, never as model input.

For factual attention overlap, the final Slot Attention allocation maps supply w_i to Eq. D.3, and the metric is averaged over every frame of all 250 validation videos. STAITUS is evaluated only on its factual active slots, consistent with the state that contributes to its reconstruction. This convention is favorable to STAITUS because closed slots are excluded, whereas GeoCo-SAVi's all-slot average

also includes diffuse empty-slot maps. This row is a factual allocation diagnostic, not an alpha-overlap or editable-geometry metric.

F.4 All-Slot Editing and External Position Evaluation

Editing uses all 24 frames of all 250 validation videos and the valid non-background-object rule in Sec. E. Each factual slot is edited independently using the four translations and five-point scale grid. Both factual edit references and interventions use the same decoder precision. STAITUS preserves its active-slot gating. The all-object and ideal-boundary-feasible scopes use the same definitions as Obj3D, with no cross-method slot association. For transplantation, donor windows are taken from a different video in cyclic order, with the same foreground-slot rank. Editing F-scores quantify command adherence using Eq. D.5.

External position errors use all 1,543 area-qualified GT tracks. The reference frame maximizes visible GT area, with earliest-frame tie-breaking. Tracks require at least 20 reference pixels; a one-pixel eroded core, or the original mask if erosion empties it, is used for factual Hungarian association. Each association is frozen before measurement. The visible GT centroid supplies $\mathbf{c}^\star$ in Table 7. These errors describe both latent position and decoded-mask localization relative to visible objects. The per-frame editing experiment uses every eligible slot independently of this external association.

F.5 Extended Numerical Results

Tables 19–24 give the per-frame all-slot measurements and their denominators. Section G reports factual scale calibration on both datasets. Boundary feasibility is computed from factual masks and their ideal transforms; the resulting evaluation counts are reported for all methods.

Table 19: MOVi-C editing results. F1 (%) uses valid edits in each scope; interior uses factual/ideal feasibility. $E_{\Delta p}$ is a diagonal fraction and D_{app} is grid L2.

Method / scope	F_{pos}	F_{scl}	F_{app}	$E_{\Delta p}$	D_{app}
STAITUS (all)	75.96	76.86	60.37	0.02442	0.08065
GeoCo-SAVi (all)	83.01	83.11	72.04	0.01045	0.04083
STAITUS (interior)	74.25	76.28	58.33	0.01745	0.05848
GeoCo-SAVi (interior)	87.86	85.85	76.02	0.00667	0.03335

Table 20: MOVi-C five-point scale response. Slopes are means over defined curves; counts are defined/total.

Method / scope	β_r	β_A	β_r def.	β_A def.
STAITUS (all)	0.70923	1.12247	50738/51971	50964/51971
GeoCo-SAVi (all)	0.85099	1.37988	50631/51629	50916/51629
STAITUS (interior)	0.73223	1.21219	12042/12159	12056/12159
GeoCo-SAVi (interior)	0.88920	1.61780	16796/16866	16802/16866

Table 21: MOVi-C scope-specific F1 denominators and defined/total center-error counts. F1 retains all actual outputs for valid edits, including empty outputs.

Method / scope	N_{pos}	N_{scl}	N_{app}	$E_{\Delta p}$ def.	D_{app} def.
STAITUS (all)	204961	201269	44284	205891/207884	42904/45177
GeoCo-SAVi (all)	198500	189003	41954	198106/206516	43175/43950
STAITUS (interior)	72936	68173	7181	72030/72936	6764/7181
GeoCo-SAVi (interior)	113245	103197	12059	112427/113245	11839/12059

Table 22: MOVi-C conditional output diagnostics, shown as count/eligible count (%). Disappearance counts empty outputs for positive-area ideal targets.

Method / scope	Position	Scale	Transplant
STAITUS (within-canvas)	906/72936 (1.24)	792/68173 (1.16)	417/7181 (5.81)
GeoCo-SAVi (within-canvas)	818/113245 (0.72)	559/103197 (0.54)	220/12059 (1.82)
STAITUS (all)	1494/206317 (0.72)	1227/207586 (0.59)	2263/45098 (5.02)
GeoCo-SAVi (all)	4172/201674 (2.07)	901/206414 (0.44)	760/43820 (1.73)

Table 23: MOVi-C pooled cross-appearance fixed-geometry statistics. Radius CV is $\sigma_{r|g}/\bar{r}$. GeoCo-SAVi has lower radius standard deviation and CV at all three scales. N counts all measurements.

Method	s	$\bar{r}$	$\sigma_{r\|g}$	$\mathrm{CV}_{r\|g}$	$\sigma_{A\|g}$	N
STAITUS	0.1	0.17068	0.16820	0.98550	0.11655	51962
STAITUS	0.2	0.24798	0.17560	0.70813	0.14393	51962
STAITUS	0.3	0.31203	0.18789	0.60217	0.17095	51962
GeoCo-SAVi	0.1	0.08325	0.02485	0.29851	0.00625	51600
GeoCo-SAVi	0.2	0.15475	0.05029	0.32496	0.01908	51600
GeoCo-SAVi	0.3	0.22067	0.08241	0.37347	0.04146	51600

Table 24: MOVi-C no-op F1 (%): unchanged recipient masks scored against the requested edit targets, on the same valid edits without boundary filtering.

Method	$F^{\mathrm{noop}}_{\mathrm{pos}}$	$F^{\mathrm{noop}}_{\mathrm{scl}}$	$F^{\mathrm{noop}}_{\mathrm{app}}$
STAITUS	48.06	64.07	59.21
GeoCo-SAVi	29.48	65.69	60.18

F.6 Whole-Video Editing

A separate experiment follows one deterministically selected GT track per video. Hungarian association at its maximum-visible-area frame fixes a slot index, which is then edited throughout all 24 recurrent states. The selected object must be valid at the reference frame, with the support-area criterion scaled to the 128×128 measurement resolution. Both methods have 247 eligible videos; 238 have the same eligible recipient and donor video pair across methods. Table 25 reports this common set. The donor is the next eligible video in cyclic order, and its appearance stream $\mathbf{a}_t$ is transplanted into the recipient's geometry stream $(\mathbf{p}_t, s_t)$.

The four translation commands and four nonidentity scale commands are held constant through the video. Targets are ideal transformations of each frame's factual support; transplantation transforms the donor's support using the corresponding frame's geometry. GT anchors the slot association at the reference frame. All 24 frames contribute, including empty edited masks and occluded frames. The tube score follows Eq. D.8 with $T = 24$. Valid-edit target area is measured over the full tube using the same resolution-scaled criterion. Scores are averaged equally over videos and then over commands within each operation. The no-op scores the unchanged recipient sequence against the requested edit target on the same population.

F.7 Object-Matched Editing and RGB Preservation

We compare edits of common GT objects using the reference-frame associations in Sec. F. The four translations and four nonidentity scales follow Sec. E.2. Each measure is averaged within source videos and then equally over videos. Table 26 reports the three evaluations below.

Common-object editing. The common pool requires valid reference edits and valid edits for both associated decoded supports, as in Sec. E.6. The support-area criterion is scaled to the 128×128

Table 25: Whole-video tube F1 (%) on 238 common recipient–donor video pairs, with all 24 frames retained. Brackets give 95% confidence intervals for the main-table F-scores using the resampling procedure in Sec. E.5.

Method	F_{pos}	F_{scl}	F_{app}
STAITUS	75.34	72.14	45.15
	[73.46, 77.12]	[70.14, 74.01]	[39.71, 50.41]
GeoCo-SAVi	86.51	84.93	74.01
	[85.68, 87.31]	[84.03, 85.82]	[70.48, 77.26]
STAITUS no-op	54.60	61.77	47.20
GeoCo-SAVi no-op	40.99	64.60	58.39

measurement resolution. Without boundary filtering, the pool contains 5,801 translation and 5,591 scale edits, each spanning 1,496 GT objects and all 250 videos. Editing F1 follows Eq. D.5.

Independent RGB targets. We construct visible-layer RGB oracle targets using official GT masks and the compositing procedure in Sec. E.6. GT masks are area-resized to 64×64 mattes, and scaling is centered at the GT centroid. The 1,541 valid source objects across 250 videos give 6,164 edits per command family. Both methods use identical targets. PSNR is measured over the full image and the source/target support union expanded by two four-connected pixel steps. Transplantation uses the reference-frame associations above and the donor-pairing and visible-layer procedure in Sec. E.6, retaining all 1,541 valid reference pairs without boundary filtering.

Non-target RGB preservation. We evaluate 178 common factual associations from 125 videos, with reference-core recall and purity at least 0.5 for both methods. All eight position and scale commands are retained, giving 712 edits per family. For these commands, the measurement region excludes the factual/ideal GT support union, nearest-resized to 64×64 and dilated with a 5×5 square. Applying the same association criterion to both members of each transplant pair retains 14 pairs from 14 recipient videos; changes are measured outside the transplant editing region defined above. We report RGB MAE and the percentage of pixels whose channel-averaged absolute change exceeds $2/255$, as in Sec. E.6. Paired 95% confidence intervals from 5,000 source-video bootstrap replicates, with fixed donor assignments for transplantation, exclude zero for all reported improvements in Table 26.

Table 26: MOVi-C object-matched editing and RGB preservation. F1 and changed-pixel rates are percentages; oracle PSNR is in dB. Each block uses common object–command pairs across methods.

EDITING EVALUATION

	F1 (%)↑		Oracle PSNR: full image↑			Oracle PSNR: editing region↑		
Method	F_{pos}	F_{scl}	Position	Scale	Transplant	Position	Scale	Transplant
STAITUS	77.03	73.50	20.444	20.587	19.000	17.202	17.796	15.234
GeoCo-SAVi	**85.37**	**84.65**	**21.929**	**21.983**	**19.932**	**18.607**	**19.145**	**15.887**

NON-TARGET RGB PRESERVATION

	RGB MAE $\times 10^3$ ↓			Changed pixels (%)↓		
Method	Position	Scale	Transplant	Position	Scale	Transplant
STAITUS	1.9314	2.5208	11.1853	4.3568	6.0451	20.4504
GeoCo-SAVi	**0.5603**	**0.7574**	**0.8834**	**1.2508**	**1.8191**	**2.2499**

G FACTUAL SCALE CALIBRATION

We assess the relationship between inferred scale s and factual decoded hard-support RMS radius r over valid non-background objects on the native canvas. Each frame–slot observation is counted once, without boundary filtering or donor–recipient weighting. We fit $r = as + b$ by least squares and $\log r = \log s + b$ with slope fixed to one. For each fit, we report $R^2 = 1 - \sum_i (y_i - \widehat{y}_i)^2 / \sum_i (y_i - \overline{y})^2$

in its respective response space (r or $\log r$). The standard deviation of $\log(r/s)$ summarizes variation in the radius-to-scale ratio. These statistics describe factual scale calibration; editing responses are evaluated separately in the preceding sections.

Table 27: Factual scale calibration on Obj3D and MOVi-C. N counts valid factual frame–slot observations. The two R^2 columns use their respective response spaces.

Dataset	Method	N	Linear $R^2 \uparrow$	Unit-slope log $R^2 \uparrow$	$\mathrm{Std}[\log(r/s)] \downarrow$
Obj3D	ISA	9424	0.6637	0.5078	0.2093
	SlotAug	9357	**0.8129**	0.6399	0.1551
	GeoCo-SAVi	9320	0.7924	**0.6873**	**0.1394**
MOVi-C	STAITUS	51971	0.6444	0.6346	0.3164
	GeoCo-SAVi	51629	**0.8589**	**0.8346**	**0.2041**

H Finite-Grid Covariance Diagnostics

We evaluate the per-slot raw-logit covariance in Proposition 1 using native-grid interior measurements and an extended-canvas diagnostic. Each object receives four translations and four nonidentity scales. With edited logits ℓ_e, bilinearly inverse-warped reference logits ℓ_t, and measurement domain D, the relative error is

$$\epsilon_\ell^{\mathrm{rel}} = \frac{\sqrt{|D|^{-1} \sum_{\mathbf{u}\in D} (\ell_e(\mathbf{u}) - \ell_t(\mathbf{u}))^2}}{\max\{\sqrt{|D|^{-1} \sum_{\mathbf{u}\in D} \ell_t(\mathbf{u})^2}, 10^{-8}\}}.$$

Statistics are averaged over interventions with at least 16 measurement pixels, using FP32 with TF32 disabled. Both diagnostics use the same frozen slots and model weights.

This diagnostic uses a separate factual selection. Obj3D selects one highest-confidence qualified slot per window, giving 2,581 objects from 2,592 evaluated windows and 10,324 interventions per edit family. Qualification requires hard area between 20 pixels and 60% of the canvas, alpha peak above 0.3, scale below 0.3, and an available background slot. Normalized attention has effective area between 2 and 75% of the tokens, maximum cosine with another slot below 0.5, and a 2.5-standard-deviation extent around the explicit position within a two-pixel canvas margin. Among qualified slots, confidence is one minus the maximum attention cosine. MOVi-C uses 178 common GT-associated objects with reference-core recall and purity at least 0.5, an available background, and a two-pixel reference bounding-box margin. These populations characterize numerical covariance on localized factual objects; all-slot editing results above use their own stated populations.

Native output, receptive-field-valid domain. We propagate validity through each bilinear upsampling and scale-steered convolution, including the alpha head. The measurement domain contains pixels for which the edited field and all bilinear contributors to the warped reference have complete in-canvas computational support (validity at least $1 - 10^{-6}$). Native rendered fields are unchanged. Domain membership depends on geometry and operator support.

Extended computational canvas, original measurement domain. We extend the per-layer coordinate domain from $[-1, 1]^2$ to $[-3, 3]^2$, preserving lattice spacing, weights, coordinate functions, and kernel offsets. A layer with native side length H thus uses $3(H - 1) + 1$ samples. Factual and edited fields are computed on this extended canvas and measured at original-grid pixels whose inverse-transformed coordinates lie in the original factual canvas. This computational diagnostic uses no receptive-field restriction on the measurement pixels.

Table 28 reports both raw-logit diagnostics on Obj3D and MOVi-C. Translation errors are below 0.06% in both settings; scaling errors range from 2.04–2.45% on the native receptive-field-valid domain and 0.92–1.42% with the extended computational canvas. These pixelwise measurement domains are separate from the object-level boundary scopes used for editing metrics.

Table 28: Finite-grid covariance and final-output diagnostics. Raw-logit columns report relative RMSE in percent, with the number of interventions having at least 16 measurement pixels in parentheses. Native uses the receptive-field-valid domain; extended uses the extended computational canvas. Final-output columns use the extended canvas: probability RMSE averages all 10,324 interventions per family on Obj3D and 712 on MOVi-C; F1 and its lower bound are means in percent over the N_{F1} target-membership comparisons with $|A| + |I| > 0$.

		Raw-logit RMSE (%), N		Extended-canvas final outputs			
Dataset	Edit	Native	Extended	Prob. RMSE	F1	F1 bound	N_{F1}
Obj3D	Translation	0.0325 (10324)	0.0199 (10324)	0.00422	95.87	95.56	10324
	Scaling	2.0376 (10324)	0.9230 (10324)	0.01049	95.69	95.17	10324
MOVi-C	Translation	0.0515 (704)	0.0446 (712)	0.00042	99.86	99.79	711
	Scaling	2.4478 (689)	1.4237 (712)	0.01136	96.95	95.16	712

H.1 Final Probabilities and Target Membership

Let ℓ_j^f and ℓ_j^e denote factual and edited final mask logits, and let P bilinearly sample a factual field at the inverse edit coordinate. For target slot i, define $q_i = P\ell_i^f$ and $\varepsilon_i = \ell_i^e - q_i$. The reference scene has target logit q_i and edited competing logits ℓ_j^e, $j \neq i$. Writing $b_e = \mathrm{logsumexp}_{j\neq i}\, \ell_j^e$ and $m_e = \max_{j\neq i} \ell_j^e$, its target probability is $\bar{\alpha}_i = \sigma(q_i - b_e)$, where σ is the logistic sigmoid.

Proposition 2 (Finite-grid probability and membership bounds)**.** *At each measurement pixel, the final full-softmax probability satisfies*

$$|\alpha_i^e - \bar{\alpha}_i| \leq \tanh\left(\frac{|\varepsilon_i|}{4}\right) \leq \frac{|\varepsilon_i|}{4}. \tag{H.1}$$

The actual and reference scenes assign the same target membership whenever

$$|q_i - m_e| > |\varepsilon_i|. \tag{H.2}$$

Let A and I be their respective argmax target masks on domain D, and let $U = \{\mathbf{u} \in D : |q_i - m_e| \leq |\varepsilon_i|\}$. For $|A| + |I| > 0$, their overlap satisfies

$$\mathrm{F1}(A, I) \geq \max\left\{0, 1 - \frac{|U \cap (A \cup I)|}{|A| + |I|}\right\}. \tag{H.3}$$

Proof. Full softmax gives $\alpha_i^e = \sigma(q_i - b_e + \varepsilon_i)$. For any displacement h, the maximum of $|\sigma(x + h) - \sigma(x)|$ over x is $\tanh(|h|/4)$; the derivative bound $\sigma'(x) \leq 1/4$ gives the second inequality. Condition (H.2) preserves the sign of the target logit's difference from the largest competitor. Thus $A \triangle I \subseteq U$, and $\mathrm{F1}(A, I) = 1 - |A \triangle I|/(|A| + |I|)$ yields Eq. (H.3). □

The residual includes the complete final-logit path. In particular, if $\ell_i = r_i + c_i$ separates raw logits and final corrections, then $\varepsilon_i = (r_i^e - Pr_i^f) + (c_i^e - Pc_i^f)$. For comparison with a transported factual probability, define $b_f = \mathrm{logsumexp}_{j\neq i} \ell_j^f$, $\Delta b = b_e - Pb_f$, and $J_i = |\sigma(q_i - Pb_f) - P\alpha_i^f|$. The same argument and the triangle inequality give

$$|\alpha_i^e - P\alpha_i^f| \leq \tanh\left(\frac{|\varepsilon_i - \Delta b|}{4}\right) + J_i, \tag{H.4}$$

which explicitly accounts for scene competition and bilinear interpolation.

Measurements. We evaluate these relations on the same objects and commands using the extended computational canvas and original measurement domain above. Final probabilities include the Obj3D selector, evaluated on the original canvas with the corresponding cropped decoder features and attention. Table 28 also reports mean probability RMSE, target-membership F1, and the mean per-intervention lower bound from Eq. (H.3). The probability inequalities hold to numerical precision, and all pixels satisfying Eq. (H.2) retain target membership.