EditCLEVR: A Paired-Scene Intervention Benchmark for Compositional Faithfulness of Object-Centric Representations

Anonymous Authors[1]

Abstract

Object-centric learning makes a concrete structural prediction: when one object changes one attribute, the corresponding object code should move, the other object codes should remain stable, and the decoded scene graph should update only at that site. Existing evaluations usually report segmentation, single-image factor prediction, or downstream accuracy, so this prediction is rarely tested as an intervention claim. We introduce **EditCLEVR**, a paired-scene benchmark in which each example contains a before/after CLEVR-style scene pair with the same layout and exactly one known attribute change on one known object, or a no-edit re-render for drift measurement. The protocol separates representation-level diagnostics for localization and stability from semantic faithfulness metrics that check whether decoded scene changes match the intended intervention across in-distribution and compositional out-of-distribution (OOD) suites. Scene-Graph Intervention Accuracy (SGIA) is the headline semantic metric: it requires the after-scene prediction to be correct and the only predicted before-to-after semantic change to be the intended object-factor edit; ΔSGIA relaxes this by checking the single-site change pattern without requiring the full after-scene graph to be correct. Across ground-truth-mask backbones, learned-slot models, SAM 2 + frozen-ViT models, and one mask-feature hybrid, EditCLEVR shows that OOD degradation persists with perfect masks, mask source explains part but not all of native performance, and locality or stability alone can overstate semantic faithfulness.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Object-centric learning is often justified by a simple structural claim: visual scenes can be represented through objects and attributes, and familiar factors should recombine in new scenes (Lake et al., 2017; Greff et al., 2020; Dittadi et al., 2022). The claim becomes sharper under intervention. If one object changes color, material, size, or shape, the matching object code should change, while other objects and other factors should stay put. Standard evaluations only test pieces of this claim. Segmentation metrics ask whether objects are found, factor probes ask whether attributes are decodable in one image, and downstream accuracy can hide shortcuts. None of them checks whether the representation carries the right before/after semantic update.

EditCLEVR turns that structural claim into a paired-scene test (Figure 1). Each evaluation unit contains two CLEVR-style renders of the same layout, instance masks, object attributes, and metadata specifying the edited object and factor. Because the intervention is known, the benchmark asks three row-level questions: did representation movement localize to the edited object, did the unedited objects remain stable, and did the decoded scene graph change only at the intended object-factor site? The paired design makes semantic faithfulness directly measurable, rather than inferred from a single-image score.

The benchmark separates ordinary in-distribution generalization from compositional out-of-distribution (OOD) transfer. Its suites cover atomic edits under the training regime, no-edit re-renders, hard-distractor cases with a similar object, and a CoGenT-style OOD split in which cube/cylinder color palettes swap between train and test (Johnson et al., 2017). A derived OOD-core slice keeps only color or shape edits on cubes and cylinders, excluding sphere cases whose color palette is unrestricted in both conditions and can dilute the intended color–shape shift.

The metric protocol is organized around design roles, not a single leaderboard number. Localization and no-edit drift measure movement directly in object-code space. Target-factor accuracy, non-target preservation, and unedited-object preservation test semantic components. SGIA then conjoins absolute after-scene correctness with the requirement that

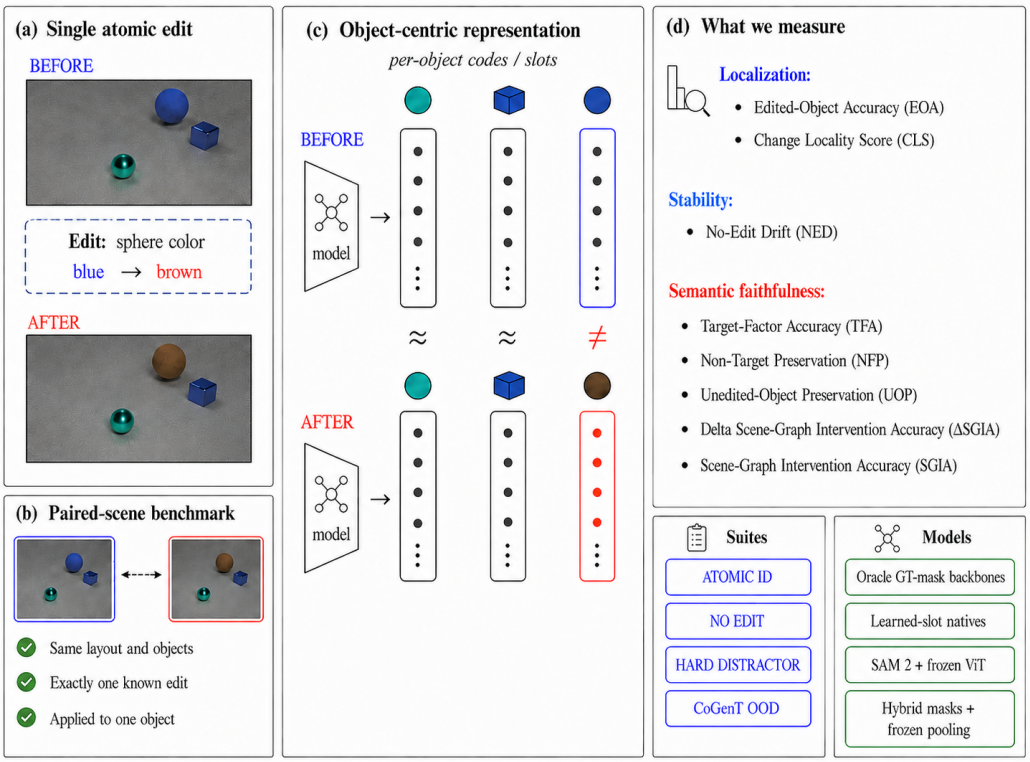

Figure 1. Evaluation protocol. EditCLEVR pairs a before scene with either a single known atomic edit or a no-edit re-render, extracts per-object representations, and evaluates whether the change localizes to the edited object, leaves other objects stable, and decodes to the intended semantic update.

the only predicted before-to-after change is the intended object-factor edit. ΔSGIA removes the absolute after-scene correctness requirement, making it a relaxed intervention-consistency diagnostic rather than a substitute for SGIA.

Contributions. EditCLEVR makes three contributions. First, it provides paired CLEVR-style scenes with known object attributes, instance masks, and either one object-factor edit or a no-edit re-render. Second, it defines a metric protocol that separates representation-level localization/stability from probe-decoded semantic faithfulness, with SGIA as the strict scene-graph measure. Third, it gives a baseline study that turns several object-centric claims into falsifiable checks: compositional OOD failures persist with ground-truth masks, and locality or stability alone can overstate semantic faithfulness.

Related work. EditCLEVR sits between object-centric evaluation and compositional generalization. Existing object-centric benchmarks commonly report discovery quality on CLEVR, Multi-dSprites, or MOVi-style scenes, or evaluate downstream factor prediction after object discovery. Disentanglement-style evaluations also measure factor decodability in single images (Higgins et al., 2017; Eastwood & Williams, 2018), but they do not test whether a decoded change follows a known intervention. Prior work shows that object-centric models can improve some forms of compositional generalization while still failing under novel factor combinations (Dittadi et al., 2022; Montero et al., 2022). EditCLEVR differs by making a before/after intervention the evaluation unit. The compared models span Slot Attention (Locatello et al., 2020), DINOSAUR-style slots over frozen ViT features (Seitzer et al., 2023), and SAM/SAM 2 segmentation-prior pipelines (Kirillov et al., 2023; Ravi et al., 2024); CLEVR-CoGenT motivates the color–shape OOD shift (Johnson et al., 2017; Didolkar et al., 2024).

2. The EditCLEVR Benchmark

Pairs and ground truth. EditCLEVR evaluates pairs rather than isolated images. Each example is a pair (I, I') of RGB renders sharing a scene layout with 3–6 CLEVR-style objects under visibility and overlap constraints. The pair is accompanied by before/after instance masks, object attributes (color, material, size, and shape), and edit metadata identifying the edited object $j^\star$ and edited factor. In edit suites, exactly one attribute of exactly one object changes. In the no-edit suite, the after image is a re-render of the same semantics with a different random seed, so movement in representation space is treated as drift rather than an edit.

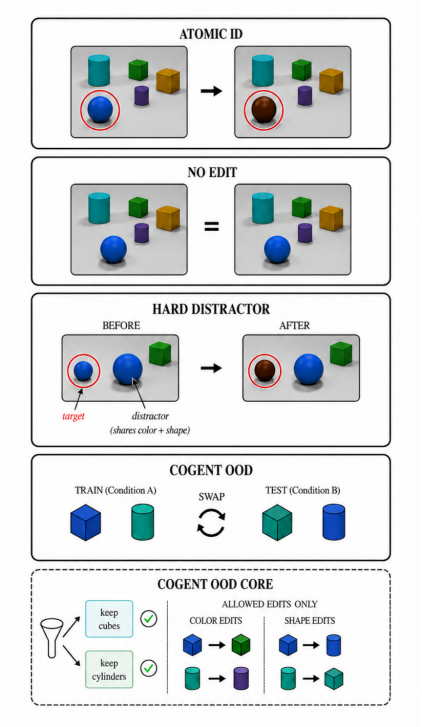

Figure 2. EditCLEVR suite design. The suites stress in-distribution semantic edits, no-edit stability, instance disambiguation, and compositional OOD transfer. The OOD-core slice removes sphere cases because spheres use the full color palette in both CoGenT conditions.

Splits and suites. The release contains 20,000 pairs: `train` (10k) and `val` (1k) for probe fitting, `test_id` (3k), `test_noop` (2k), `test_hard` (2k), and `test_cogent` (2k). These form four suites: ATOMIC_ID tests condition-A edits, NO_EDIT measures drift under semantic no-ops, HARD_DISTRACTOR tests target/distractor separation, and COGENT_OOD tests condition-B color–shape transfer (Figure 2). The derived COGENT_OOD_CORE slice keeps only color or shape edits on cubes or cylinders before and after the edit, excluding sphere cases because spheres use the full color palette in both CoGenT conditions. Edit factors and object counts are balanced across splits.

3. Metrics

For each pair, let T be the trusted object set, $j^\star$ the edited object, $f^\star$ the edited factor, z_j, z'_j the L^2-normalized object vectors, and $d_j = \|z_j - z'_j\|_2$. Oracle rows use all objects; native rows use the assignment and overlap gates in Section 4. Localization and stability are defined directly in object-code space: $\mathrm{EOA} = \mathbf{1}[\arg\max_{j\in T} d_j = j^\star]$, $\mathrm{CLS} = d_{j^\star} / \sum_{j\in T} d_j$, and $\mathrm{NED} = |T|^{-1} \sum_{j\in T} d_j$ on no-edit re-renders.

For semantic metrics, let $y_{j,f}, y'_{j,f}$ be ground-truth attributes and $\hat{y}_{j,f}, \hat{y}'_{j,f}$ the probe predictions. TFA requires $\hat{y}'_{j^\star,f^\star} = y'_{j^\star,f^\star}$; NFP requires $\hat{y}'_{j^\star,f} = \hat{y}_{j^\star,f}$ for all $f \neq f^\star$; UOP requires $\hat{y}'_{j,f} = \hat{y}_{j,f}$ for all non-edited $j \in T$. The single-site condition Δ_{site} holds when the only predicted before-to-after change is $(j^\star, f^\star)$. We set $\Delta\text{SGIA} = \text{TFA} \wedge \Delta_{\text{site}}$, while $\text{SGIA} = \text{SceneGraphExact} \wedge \Delta_{\text{site}}$, where SceneGraphExact requires every decoded after-frame attribute in T to equal ground truth. Thus ΔSGIA tests intervention consistency, whereas SGIA also requires full after-scene correctness. Metrics are averaged over rows with 95% bootstrap confidence intervals.

4. Models and Evaluation Protocol

Model families as controls. We vary how object regions are obtained and which features are pooled. Ground-truth-mask rows pool frozen patch tokens under true instance masks, isolating representation quality when segmentation is perfect. These rows use DINO ViT-S/8 (384-d, 224 px) (Caron et al., 2021), DINOv2 ViT-B/14 (768-d) (Oquab et al., 2023), and SigLIP 2 ViT-B/16 (768-d, 384 px) (Tschannen et al., 2025).

Native object discovery. The learned-slot rows test models that must discover objects: convolutional Slot Attention (SA) (Locatello et al., 2020) and DINOSAUR, which applies Slot Attention to frozen DINO ViT-S/8 patch tokens with an MLP patch decoder (Seitzer et al., 2023). The SAM 2 rows use automatic proposals (Kirillov et al., 2023; Ravi et al., 2024) and pool frozen ViT patch features inside those masks. The hybrid row uses DINOSAUR masks but frozen DINO-S/8 pooled features, giving a controlled mask-source comparison within one backbone family.

Native matching and gates. For native rows, we match predicted objects to ground-truth instances separately in each frame with a strict one-to-one best-overlap assignment; unused slots and extra SAM 2 proposals are ignored. Semantic metrics require the edited object to have MatchBO ≥ 0.5 in both frames, and UOP/SceneGraphExact use objects assigned in both frames. Native rows also report FG-ARI, MatchBO, MatchIoU, and low-confidence match rate. A soft IoU-mixture alternative appears in Appendix D.

5. Results

Table 1 supports three observations. First, ID–OOD degradation persists even with ground-truth masks: SGIA drops from 0.866 to 0.459 for SigLIP 2, 0.821 to 0.190 for DINO, and 0.691 to 0.067 for DINOv2; the corresponding ΔSGIA drops are 22%, 55%, and 74%. Since these rows use true masks, the gap is not a discovery failure. It is consistent with limited transfer of per-object features and probes trained on condition-A combinations, although this exper-

Table 1. Linear-probe results on EditCLEVR. ID is ATOMIC_ID ($n = 3{,}000$); OOD is COGENT_OOD_CORE ($n = 327$); NED uses NO_EDIT ($n = 2{,}000$). Values are row-level means; Table 3 reports 95% bootstrap CIs for SGIA and ΔSGIA. Higher is better except NED. Best overall values are **bold**; best native values are underlined.

	ATOMIC_ID (ID)				COGENT_OOD_CORE (OOD)					
Model	SGIA	ΔSGIA	EOA	CLS	SGIA	ΔSGIA	EOA	TFA	NED ↓	ΔSGIA gap
Ground-truth-mask frozen backbones										
DINO ViT-S/8	0.821	0.877	0.946	0.517	0.190	0.391	0.914	0.694	0.138	−0.486
DINOv2 ViT-B/14	0.691	0.799	0.930	0.476	0.067	0.211	0.847	0.321	0.135	−0.588
SigLIP 2 ViT-B/16	**0.866**	**0.911**	0.947	0.509	**0.459**	**0.710**	**0.960**	**0.976**	0.181	−0.201
Native (learned-slot) discovery										
SA (conv)	0.022	0.149	0.584	0.454	0.007	0.229	0.927	0.980	0.140	+0.080
DINOSAUR	0.513	0.619	0.763	0.410	0.083	0.269	0.859	0.503	0.111	−0.350
Native (SAM 2 proposals + frozen ViT)										
SAM 2 + DINO-S/8	0.482	0.589	0.761	0.464	0.199	0.347	0.810	0.742	0.165	−0.242
SAM 2 + DINOv2	0.446	0.550	0.789	0.433	0.095	0.209	0.746	0.439	0.163	−0.341
SAM 2 + SigLIP 2	0.514	0.615	0.785	0.464	0.098	0.344	0.872	0.942	0.198	−0.271
Hybrid (predicted masks + frozen pooling)										
DINOSAUR-mask + DINO-S/8	0.175	0.323	0.480	0.265	0.046	0.173	0.462	0.568	**0.047**	−0.150

iment does not isolate pretraining objective, data, resolution, or probe form. Hard-distractor results mostly track ATOMIC_ID (Appendix B), so we treat that suite as an instance-disambiguation diagnostic rather than the main stress test.

Second, mask geometry explains part, but not all, of native-model performance. Holding the feature family fixed at DINO ViT-S/8, ground-truth masks give ID ΔSGIA 0.877, SAM 2 masks give 0.589, and DINOSAUR masks with frozen DINO pooling give 0.323. Table 6 fits this ordering: SAM 2 has higher MatchIoU than DINOSAUR, while DINOSAUR has high MatchBO but low IoU. Still, masks are not the whole story. DINOSAUR's learned slot features reach ID ΔSGIA 0.619 with the same mask family.

Third, locality and stability do not guarantee semantic faithfulness. The DINOSAUR-mask hybrid has the lowest no-edit drift (NED 0.047) but low SGIA (ID 0.175, OOD-core 0.046). Conversely, SA reaches TFA 0.980 on OOD-core but SGIA 0.007. A row is counted as faithful only when the edited factor, non-target factors, unedited objects, and full after-scene graph are all correct; Appendix F details the failure accounting.

6. Discussion and Conclusion

EditCLEVR is meant as a small benchmark for a specific theoretical claim: object-centric representations should support localized, stable, semantically correct interventions. The results show why that claim needs a paired test. A model can discover objects, decode the edited factor, or remain stable under re-rendering noise while still changing the wrong parts of the predicted scene graph. SGIA rewards only rows where localization, preservation, and after-scene correctness all hold. The results point to two interacting bottlenecks: ground-truth masks do not remove the CoGenT-style ID–OOD gap, and mask source affects native performance without fully explaining it. SigLIP 2 is strongest on OOD-core ΔSGIA and DINOv2 is most brittle, which raises a pretraining hypothesis; the present study does not isolate objective or data from architecture and resolution.

Limitations and future work. EditCLEVR is synthetic and restricted to CLEVR-style objects, four discrete attributes, and single-object edits. The baselines cover ground-truth masks, learned slots, SAM 2 proposals, frozen ViTs, and one hybrid, but not all object-centric or generative representation learners. The semantic protocol uses supervised probes; the OOD drop persists under a two-layer MLP probe (Appendix C), but probe design still affects rankings. Natural images, relational or continuous edits, simultaneous edits, stronger discovery sources, calibration baselines, and factor-structured readouts are the next tests. We will release the paired scenes, masks, edit metadata, and metric pipeline at the camera-ready stage; Appendix G lists the planned artifacts.

Impact Statement

EditCLEVR is a synthetic diagnostic benchmark. Its direct societal impact is limited, and it should not be read as evidence of real-world visual robustness; controlled scenes and supervised probes complement, but do not replace, natural-image and safety-critical evaluations.

References

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Didolkar, A., Zadaianchuk, A., Goyal, A., Mozer, M. C., Bengio, Y., Martius, G., and Seitzer, M. Zero-shot object-centric representation learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Dittadi, A., Papa, S. S., De Vita, M., Schölkopf, B., Winther, O., and Locatello, F. Generalization and robustness implications in object-centric learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5221–5285. PMLR, 2022.

Eastwood, C. and Williams, C. K. I. A framework for the quantitative evaluation of disentangled representations. In *Proceedings of the 6th International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=By-7dz-AZ.

Greff, K., van Steenkiste, S., and Schmidhuber, J. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017. doi: 10.1109/CVPR.2017.215.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538, 2020.

Montero, M. L., Bowers, J. S., Ponte Costa, R., Ludwig, C. J. H., and Malhotra, G. Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://openreview.net/forum?id=7yUxTNWyQGf.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=a68SUt6zFt.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C.-J., He, T., Zhang, Z., Schölkopf, B., Brox, T., and Locatello, F. Bridging the gap to real-world object-centric learning. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=b9tUk-f_aG.

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

A. Dataset and Implementation Details

Dataset. We render 20,000 paired scenes with a CLEVR-derived generator: $3-6$ objects per scene, balanced over the four edit factors, and constrained by visibility and overlap floors. Splits are `train`/`val`/`test_id` (10k/1k/3k, suite ATOMIC_ID), `test_noop` (2k, NO_EDIT), `test_hard` (2k, HARD_DISTRACTOR), and `test_cogent` (2k, COGENT_OOD). The CoGenT regime swaps cube/cylinder color palettes between condition A (train/val/test_id/test_noop/test_hard) and condition B (test_cogent). The COGENT_OOD_CORE slice keeps rows whose edit factor is color or shape and whose edited object is a cube or cylinder both before and after the edit (327 rows). For each pair we ship before/after RGB renders, before/after instance masks, before/after attribute records, and edit metadata.

Backbones and discovery. DINO ViT-S/8 and DINOv2 ViT-B/14 run at 224×224; SigLIP 2 ViT-B/16 runs at 384×384. Slot Attention (Locatello et al., 2020) (denoted SA) uses a convolutional encoder at the dataset resolution with the slot count set above the maximum object count of the data; evaluation uses the same one-to-one best-overlap assignment as the other native rows. DINOSAUR (Seitzer et al., 2023) pools 28×28 frozen DINO ViT-S/8 patch tokens through Slot Attention with an MLP patch decoder; we export both predicted slot masks (for native rows and the hybrid) and slot features. SAM 2 (Ravi et al., 2024) runs in automatic-mask mode on each frame; per-mask proposals are then pooled with frozen ViT patch tokens and L_2-normalized.

Object vectors and matching. All object vectors are L_2-normalized in $\mathbb{R}^d$ (oracle pooled patch tokens for oracles; slot features for slot natives; mask-pooled patch tokens for SAM 2 and the hybrid). Native rows use the strict one-to-one best-overlap assignment described in Section 4; each predicted object can be assigned to at most one ground-truth instance within a frame. Semantic metrics use a $\mathrm{MatchBO}\geq0.5$ gate on the edited object in both frames. The soft IoU mixture in Appendix D is an ablation, not the headline protocol.

Probes. The primary probe is one LOGISTICREGRESSION per factor (color, material, size, shape) trained on `train` object vectors with seed-averaged accuracy. The MLP replication uses a 2-layer MLPCLASSIFIER (512×512, Adam, early stopping, seed 42). All metrics are computed per row; CIs are 95% bootstrap intervals over per-row arrays.

B. Full Per-Suite Linear Results

Table 2 reports SGIA / ΔSGIA across all four suites for every model. SGIA drops from ATOMIC_ID to COGENT_OOD_CORE for every model. ΔSGIA drops for every model except SA, which rises ($0.149\rightarrow0.229$); however, SA's SGIA on both suites is essentially zero, so the apparent ΔSGIA gain reflects single-site change patterns under near-saturated NFP/UOP failure rather than improved intervention faithfulness.

Table 2. Linear-probe SGIA / ΔSGIA across all four suites. NED is on NO_EDIT.

	ATOMIC_ID		HARD_DISTRACTOR		COGENT_OOD		COGENT_OOD_CORE		
Model	SGIA	ΔSGIA	SGIA	ΔSGIA	SGIA	ΔSGIA	SGIA	ΔSGIA	NED ↓
DINO ViT-S/8 (oracle)	0.821	0.877	0.821	0.879	0.161	0.518	0.190	0.391	0.138
DINOv2 ViT-B/14 (oracle)	0.691	0.799	0.679	0.794	0.037	0.372	0.067	0.211	0.135
SigLIP 2 ViT-B/16 (oracle)	0.866	0.911	0.878	0.915	0.400	0.670	0.459	0.710	0.181
SA (conv, native)	0.022	0.149	0.023	0.150	0.007	0.122	0.007	0.229	0.140
DINOSAUR (native)	0.513	0.619	0.509	0.609	0.056	0.339	0.083	0.269	0.111
SAM 2 + DINO-S/8	0.482	0.589	0.472	0.572	0.154	0.383	0.199	0.347	0.165
SAM 2 + DINOv2	0.446	0.550	0.427	0.532	0.062	0.321	0.095	0.209	0.163
SAM 2 + SigLIP 2	0.514	0.615	0.515	0.602	0.102	0.378	0.098	0.344	0.198
DINOSAUR-mask + DINO-S/8	0.175	0.323	0.189	0.340	0.024	0.164	0.046	0.173	0.047

C. MLP Probe Replication

We replace the per-factor logistic regression with a 2-layer MLP (512×512, seed 42) and re-evaluate every model on ATOMIC_ID and COGENT_OOD_CORE. Table 4 shows that the MLP raises every model's ID score (it is a strictly stronger decoder) but *does not* close the OOD chasm: SGIA on OOD-core remains low for every model (best 0.171, worst 0.000),

Table 3. Bootstrap confidence intervals for the headline semantic metrics in Table 1. Intervals are 95% bootstrap CIs over row-level arrays. ID is ATOMIC_ID; OOD is COGENT_OOD_CORE.

Model	ID SGIA	ID ΔSGIA	OOD SGIA	OOD ΔSGIA
DINO ViT-S/8	[0.806, 0.835]	[0.865, 0.889]	[0.150, 0.239]	[0.339, 0.447]
DINOv2 ViT-B/14	[0.674, 0.707]	[0.784, 0.813]	[0.043, 0.095]	[0.165, 0.257]
SigLIP 2 ViT-B/16	[0.853, 0.878]	[0.900, 0.922]	[0.401, 0.514]	[0.658, 0.755]
SA (conv)	[0.017, 0.027]	[0.137, 0.162]	[0.000, 0.017]	[0.182, 0.278]
DINOSAUR	[0.493, 0.529]	[0.601, 0.636]	[0.056, 0.114]	[0.219, 0.321]
SAM 2 + DINO-S/8	[0.464, 0.500]	[0.572, 0.607]	[0.156, 0.242]	[0.295, 0.399]
SAM 2 + DINOv2	[0.429, 0.463]	[0.533, 0.567]	[0.067, 0.129]	[0.166, 0.252]
SAM 2 + SigLIP 2	[0.497, 0.532]	[0.597, 0.632]	[0.068, 0.129]	[0.295, 0.396]
DINOSAUR-mask + DINO-S/8	[0.163, 0.188]	[0.306, 0.339]	[0.025, 0.071]	[0.133, 0.216]

and SigLIP 2's SGIA drops from 0.919 ID to 0.070 OOD (-92%). The OOD chasm is therefore not a probe-capacity artifact. However, the OOD-SGIA *ranking* of models is not preserved between linear and MLP probes (e.g., SigLIP 2 oracle moves from rank 1 under linear to rank 7 under MLP). Inspecting the per-component metrics, the MLP improves ID decoding across the board but fails to preserve the joint NFP/UOP conditions on OOD-core, which is what flips the SGIA conjunction; we leave a mechanistic study to future work and use the linear probe as the primary protocol, with the MLP reported as a robustness check.

Table 4. MLP-probe SGIA / ΔSGIA / TFA on the headline suites. ID is ATOMIC_ID, OOD is COGENT_OOD_CORE.

	ATOMIC_ID (ID)			COGENT_OOD_CORE (OOD)		
Model	SGIA	ΔSGIA	TFA	SGIA	ΔSGIA	TFA
DINO ViT-S/8 (oracle)	0.901	0.924	0.996	0.171	0.431	0.710
DINOv2 ViT-B/14 (oracle)	0.783	0.865	0.985	0.089	0.288	0.404
SigLIP 2 ViT-B/16 (oracle)	0.919	0.941	0.995	0.070	0.587	0.991
SA (conv, native)	0.333	0.408	0.908	0.000	0.758	0.980
DINOSAUR (native)	0.556	0.645	0.966	0.114	0.315	0.553
SAM 2 + DINO-S/8	0.519	0.615	0.942	0.163	0.350	0.693
SAM 2 + DINOv2	0.488	0.585	0.935	0.071	0.221	0.408
SAM 2 + SigLIP 2	0.527	0.626	0.943	0.049	0.377	0.939
DINOSAUR-mask + DINO-S/8	0.465	0.562	0.955	0.161	0.336	0.806

D. Soft-Mixture Ablation for Native Models

The strict argmax-BO assignment in the main text picks one predicted object (slot or SAM 2 proposal) per ground-truth object. As an ablation, we replace this with a soft IoU-weighted convex combination of predicted-object features per ground-truth object, $f_{\text{obj}} = \sum_k w_k \, s_k$ with $w_k \geq 0$ and $\sum_k w_k = 1$, where w_k is normalized from the per-proposal IoU against the ground-truth mask. Table 5 shows that soft mixtures help SAM 2-pooled rows substantially on ATOMIC_ID—SAM 2+DINO-S/8 SGIA jumps from strict 0.482 to soft 0.660, SAM 2+SigLIP 2 from 0.514 to 0.705, and SAM 2+DINOv2 from 0.446 to 0.599—and lift SA's EOA and OOD ΔSGIA. However, soft mixtures do *not* uniformly improve every native model: DINOSAUR slot ΔSGIA-ID drops from strict 0.619 to soft 0.599, and its OOD-core ΔSGIA also drops slightly. The strict argmax is therefore not always conservative; we keep it as the headline protocol for cross-model comparability and report soft mixtures as an ablation rather than as a uniform improvement.

E. Native Discovery Diagnostics

Table 6 reports the discovery-side diagnostics that gate the semantic metrics for native and hybrid rows: foreground ARI (FG-ARI), best-overlap (MatchBO), match IoU (MatchIoU), and low-confidence rate. DINOSAUR's predicted masks have

Table 5. Soft IoU-mixture native objectization (linear probes). Compare with the corresponding rows of Table 2 (strict argmax). NED uses the same gating.

	ATOMIC_ID		COGENT_OOD_CORE			
Model	SGIA	ΔSGIA	SGIA	ΔSGIA	EOA(ID)	NED ↓
SA (conv, native)	0.024	0.155	0.000	0.255	0.631	0.083
DINOSAUR (native)	0.483	0.599	0.080	0.259	0.819	0.090
SAM 2 + DINO-S/8	0.660	0.741	0.245	0.411	0.863	0.146
SAM 2 + DINOv2	0.599	0.690	0.101	0.252	0.872	0.148
SAM 2 + SigLIP 2	0.705	0.777	0.236	0.534	0.871	0.185

very high BO (0.96) but low IoU (0.12): the masks find each ground-truth object reliably but cover much more than the object, which is consistent with the hybrid model's residual gap to the DINO oracle (Table 1). SAM 2 sits in the middle (IoU ~ 0.29) and has the lowest low-confidence rate.

Table 6. Discovery diagnostics on ATOMIC_ID. Higher is better except low-confidence rate.

Model	FG-ARI	MatchBO	MatchIoU	LowConf ↓
SA (conv)	0.859	0.904	0.476	0.016
DINOSAUR	0.936	0.959	0.123	0.005
SAM 2 + DINO/DINOv2/SigLIP 2	0.907	0.970	0.294	0.002
DINOSAUR-mask + DINO-S/8	0.936	0.959	0.123	0.005

F. What Gets Counted as a Failure

SGIA is intentionally strict: a row passes only when the after-frame scene graph is exactly correct on all trusted objects ($\mathrm{SceneGraphExact} = 1$) *and* the before-to-after prediction change pattern is confined to the edited object's edited factor (the ΔSGIA single-site change pattern). This subsumes target-factor correctness, non-target preservation, and unedited-object preservation, but it additionally requires absolute after-scene correctness through SceneGraphExact, which is what makes it stricter than any factor-level conjunction. Each component term has a high natural baseline—e.g., TFA exceeds 0.99 for SigLIP 2 oracle on ID and remains 0.98 on OOD-core—yet the joint event fails dramatically OOD. SA is the most extreme example: its TFA on COGENT_OOD_CORE is 0.980 but its SGIA is 0.007, because the absolute after-scene predictions and the single-site change pattern almost never both hold. We therefore use SGIA, rather than single-factor decoding accuracy, as the headline number for paired-scene intervention benchmarks.

G. Reproducibility

We commit to releasing at the camera-ready stage: (i) the dataset generator (Blender pipeline plus split-validators that pin CoGenT condition A/B), (ii) all model checkpoints and SAM 2 mask caches, (iii) the per-model feature-extraction notebook, and (iv) the probe-and-metric notebook that produces every number in this paper, with seed-fixed bootstrap CIs. The release will also include the run tag and provenance metadata needed to reproduce each table.