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

Anuraag Gadehothur Karnam [1 *] **Tarunesh Sathish** [2 *]

Abstract

Object-centric learning aims to represent scenes as objects whose properties can be reused in new combinations. Existing evaluations usually score segmentation, single-image factor prediction, or downstream accuracy, but these tests do not directly ask whether a per-object representation behaves correctly under a controlled semantic edit. We introduce **EditCLEVR**, a paired-scene intervention benchmark in which each example contains a before/after pair of CLEVR-style renders with the same object indices and scene layout, and either exactly one known attribute change on one known object or a no-edit re-render for drift measurement. The protocol includes probe-free diagnostics for representation-change localization and stability, together with probe-decoded semantic faithfulness metrics that test whether the predicted scene change matches the intended intervention across in-distribution and compositional out-of-distribution (OOD) suites, allowing code-space movement and decoded object-attribute correctness to be evaluated separately. We introduce the semantic metric Scene-Graph Intervention Accuracy (SGIA), which requires the full after-scene prediction to be correct and the only predicted before-to-after semantic change to be the intended object-factor edit. We also establish ΔSGIA as a companion diagnostic that checks the single-site change pattern without requiring the full after-scene graph to be correct. Baseline evaluations on ground-truth-mask backbones, learned-slot models, SAM 2 + frozen-ViT models, and one mask-feature hybrid indicate that CoGenT-OOD-core degradation can persist under ground-truth instance masks, that mask source accounts for part but not all of native performance, and that locality or stability alone can overstate semantic faithfulness. Code is available at https://github.com/torux-bughunter/EditCLEVR.

[*]These authors contributed equally. [1]Lebanon Trail High School [2]Walnut Grove High School. Correspondence to: Anuraag Gadehothur Karnam <anuraag.gadehothurkarnam.761@k12.friscoisd.org>, Tarunesh Sathish <sathitar000@k12.prosper-isd.net>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Object-centric learning is motivated by the idea that visual scenes can be represented through objects and their attributes, so that familiar factors can be recombined in new scenes (Lake et al., 2017; Greff et al., 2020a; Dittadi et al., 2022). This notion is fundamentally an intervention-level one: if a single object changes color, material, size, or shape, the corresponding object code should change while other objects and factors remain stable. However, common evaluations only partially test this behavior. Segmentation metrics ask whether objects are discovered, factor probes ask whether attributes are decodable in a single image, and downstream accuracy mixes representation quality with task-specific shortcuts. These tests typically do not directly verify that a representation carries the intended before/after semantic change.

EditCLEVR operationalizes this criterion as a paired-scene benchmark (Figure 1). Each evaluation unit contains two CLEVR-style renders with the same scene layout and object indices, instance masks, object attributes, and metadata specifying the edited object and factor. Because the intervention is known, the benchmark can examine whether representation movement localizes to the edited object, unedited objects remain stable, and the decoded object-attribute graph changes only in the intended way. This paired design makes semantic faithfulness observable rather than inferred from a single-image score.

The benchmark stresses both standard generalization and compositional out-of-distribution (OOD) transfer. Its suites cover in-distribution atomic edits, 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

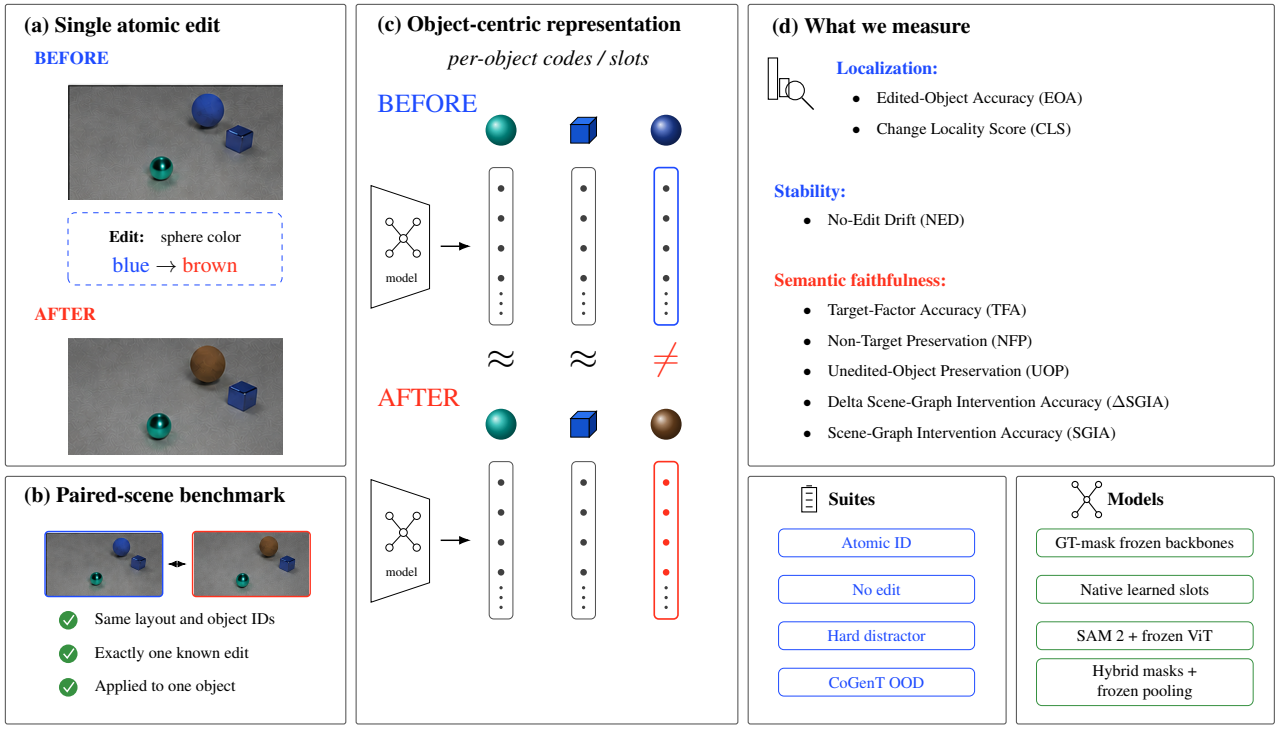

Figure 1. Evaluation protocol. EditCLEVR pairs a before scene with either a single known atomic edit or a no-edit re-render, extracts per-object representations, and evaluates whether the change localizes to the edited object, leaves other objects stable, and decodes to the intended object-attribute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

cylinders, excluding sphere cases whose color palette is unrestricted in both conditions and can dilute the intended color–shape shift.

We evaluate each intervention in terms of representation movement, no-edit stability, and decoded semantic change. Change locality (measured by CLS) and no-edit drift (NED) are probe-free diagnostics. Target-factor accuracy (TFA), non-target preservation (NFP), and unedited-object preservation (UOP) test semantic components of the edit. SGIA then conjoins absolute after-scene correctness with the requirement that the only predicted before-to-after change is the intended object-factor edit. This definition is deliberately strict, so we report it alongside a companion metric ΔSGIA, which removes the absolute after-scene correctness requirement and acts as a relaxed diagnostic that measures intervention consistency.

Contributions. EditCLEVR contributes: (i) a dataset of CLEVR-derived paired scenes with known object attributes, instance masks, and either one object-factor edit or a no-edit re-render; (ii) a metric protocol that reports representation-level localization and stability alongside probe-decoded semantic faithfulness, with SGIA as the strict object-attribute scene-graph measure and ΔSGIA/component metrics as diagnostics; and (iii) a baseline study indicating that compositional OOD degradation persists with ground-truth masks and that locality or stability alone can overstate semantic faithfulness.

Related work. EditCLEVR connects object-centric evaluation with compositional generalization. Object-centric benchmarks commonly report discovery quality on CLEVR, Multi-dSprites, or MOVi scenes (Johnson et al., 2017; Greff et al., 2020b; 2022), or evaluate downstream factor prediction after object discovery. Disentanglement evaluations also measure factor decodability in single images (Higgins et al., 2017; Eastwood & Williams, 2018), while identifiability results caution that such scores should not be read as proof that factors have been recovered without assumptions or supervision (Locatello et al., 2019). EditCLEVR therefore makes a before/after intervention the evaluation unit and treats probe-decoded semantics as an operational readout, not as an identifiability claim.

Paired before/after evaluation has also appeared in vision-language benchmarking. BEAF, for example, manipulates visual scenes and evaluates whether VLMs understand the resulting changes (Moon et al., 2024). EditCLEVR targets a different object of study: per-object visual representations. EditCLEVR extends paired before/after evaluation by making each scene pair object-indexed and measurable at three levels: code-space movement, object-discovery/matching, and decoded object-attribute graph change. It therefore

asks whether object codes localize the edited object, preserve unedited objects, and decode to a single intended object-attribute change. The compared models span Slot Attention (Locatello et al., 2020), DINOSAUR slots over frozen ViT features (Seitzer et al., 2023), and SAM/SAM 2 segmentation-prior pipelines (Kirillov et al., 2023; Ravi et al., 2025). CLEVR-CoGenT motivates the color–shape OOD shift (Johnson et al., 2017), and recent object-centric work also studies zero-shot transfer across datasets (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 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, object positions, and object attributes with a different renderer seed, so movement in representation space is treated as drift rather than an edit.

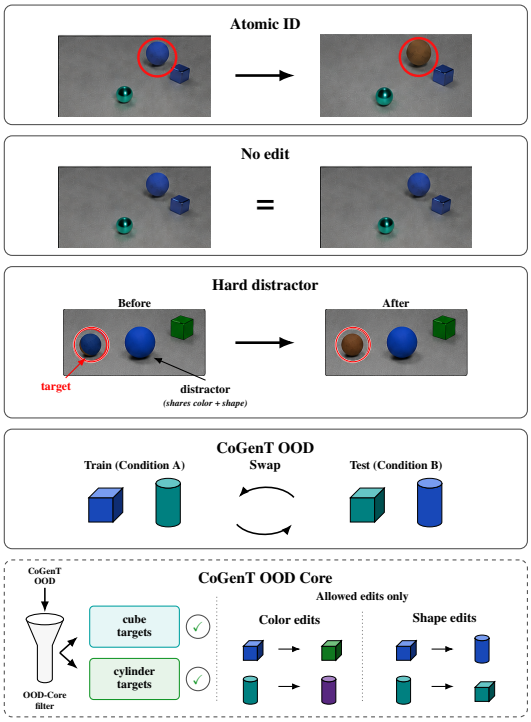

Figure 2. EditCLEVR suite design. The suites stress in-distribution semantic edits, no-edit stability, instance disambiguation, and compositional OOD transfer. The OOD-core slice restricts evaluation to color/shape edits on cube/cylinder targets and removes sphere cases because spheres use the full color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

Splits and suites. EditCLEVR contains 20,000 paired scenes. We use 10,000 training pairs and 1,000 validation pairs for probe fitting, then evaluate on four held-out suites: 3,000 atomic-ID pairs for in-distribution edits, 2,000 no-edit pairs for measuring drift under semantic no-ops, 2,000 hard-distractor pairs for target/distractor separation, and 2,000 CoGenT-OOD pairs for condition-B color-shape transfer. We also report a derived CoGenT-OOD-core slice containing only color or shape edits on cubes or cylinders before and after the edit ($n = 327$). This slice removes sphere cases, since spheres use the full color palette in both CoGenT conditions and can weaken the intended color-shape shift. The generated splits are balanced over edit factors and object counts before this OOD-core filtering step. The hard-distractor suite requires a visually similar object in the scene, but it is not meant to define a calibrated perceptual-difficulty scale; we use it as an instance-disambiguation stress test.

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 $d_j = \|z_j - z'_j\|_2$. Ground-truth-mask rows use all ground-truth objects; predicted-object rows use the assignment protocol in Section 4, with additional matching gates for semantic metrics. Localization and stability are defined directly in object-code space:

$$\mathrm{EOA} = \mathbf{1}\left[\arg\max_{j\in T} d_j = j^\star\right],$$

$$\mathrm{CLS} = \frac{d_{j^\star}}{\sum_{j\in T} d_j},$$

$$\mathrm{NED} = |T|^{-1}\sum_{j\in T} d_j \quad \text{on no-edit re-renders.}$$

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. We compute these metrics on edit suites; no-edit rows are used for NED. TFA requires the predicted after value of the edited factor to be correct,

$$\hat{y}'_{j^\star,f^\star} = y'_{j^\star,f^\star}.$$

NFP requires the edited object's non-target factors to be unchanged in prediction space, $\hat{y}'_{j^\star,f} = \hat{y}_{j^\star,f}$ for all $f \neq f^\star$. UOP requires the predictions for every non-edited object to be unchanged, $\hat{y}'_{j,f} = \hat{y}_{j,f}$ for all $j \neq j^\star$ in T. Thus TFA checks whether the intended after attribute is decoded, while NFP and UOP test whether the decoded change leaks to other factors or objects.

The single-site condition Δ_{site} holds when the edited factor prediction changes at $(j^\star, f^\star)$ and no other predicted attribute in T changes. We set

$$\Delta\mathrm{SGIA} = \mathrm{TFA} \wedge \Delta_{\mathrm{site}},$$
$$\mathrm{SGIA} = \mathrm{SceneGraphExact} \wedge \Delta_{\mathrm{site}},$$

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; higher is better except NED. Native semantic metrics are conditional on the matching gate described in Section 4. ΔSGIA gap is OOD ΔSGIA minus ID ΔSGIA. 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	0.817	0.856	**0.905**	0.386	**0.135**	0.310	**0.898**	0.787	0.240	−0.546
DINOv2 ViT-B/14	0.634	0.749	0.833	0.381	0.044	0.117	0.681	0.310	0.230	−0.632
SigLIP 2 ViT-B/16	**0.862**	**0.884**	0.841	0.377	0.123	**0.398**	0.877	**0.980**	0.270	−0.486
Native learned-slot discovery										
SA (conv)	0.011	0.125	0.573	**0.437**	0.000	0.163	0.594	0.863	0.073	+0.038
DINOSAUR	0.640	0.694	0.799	0.374	0.021	0.137	0.696	0.395	0.116	−0.556
Native SAM 2 proposals + frozen ViT										
SAM 2 + DINO-S/8	0.485	0.560	0.713	0.360	0.111	0.237	0.702	0.798	0.264	−0.324
SAM 2 + DINOv2	0.557	0.631	0.762	0.363	0.061	0.155	0.640	0.406	0.248	−0.476
SAM 2 + SigLIP 2	0.512	0.583	0.701	0.355	0.061	0.228	0.719	0.942	0.284	−0.355
Hybrid predicted masks + frozen pooling										
DINOSAUR-mask + DINO-S/8	0.112	0.210	0.331	0.236	0.009	0.035	0.333	0.547	**0.070**	−0.175

where SceneGraphExact denotes after-frame object-attribute graph exactness: every decoded after-frame attribute in T must equal ground truth, while the before frame enters through the predicted-change pattern. SGIA therefore conjoins two requirements by design–the after scene must decode correctly, and the predicted change pattern must be single-site–so it should be read together with ΔSGIA and the component metrics. Because all semantic metrics are supervised-probe readouts, they test whether the representation supports a low-capacity decoder for the intervention semantics rather than proving that the latent representation is uniquely identifiable; Appendix C reports an MLP-probe replication to check probe sensitivity, following standard probe-interpretation cautions (Hewitt & Liang, 2019).

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 masks are controlled. 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., 2025) 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. These native semantic scores are therefore conditional on matched objects, not full end-to-end discovery-plus-faithfulness scores. Appendix E reports FG-ARI, MatchBO, MatchIoU, and the low-confidence MatchBO-exclusion rate; on ATOMIC_ID, that exclusion rate is between 0.1% and 1.4% across native rows. A soft IoU-mixture alternative appears in Appendix D.

5. Results

The results in Table 1 highlight three empirical patterns. First, the ID–OOD-core drop persists in ground-truth-mask rows: DINO, DINOv2, and SigLIP 2 backbones all lose most of their strict SGIA on COGENT_OOD_CORE, so the drop is not explained solely by object discovery. Second, mask source accounts for part, but not all, of the native-row gap: within the DINO-S/8 family, ground-truth masks give ID ΔSGIA 0.856, SAM 2 masks give 0.560, and DINOSAUR masks with frozen pooling give 0.210, while DINOSAUR's learned slot features reach 0.694. Third, locality and stability can overstate semantic faithfulness: the DINOSAUR-mask hybrid has the lowest NED (0.070) but weak SGIA, and SA has OOD-core TFA 0.863 with strict

SGIA 0.000. Appendix B reports all edit suites, Appendix F gives factor/object-count views, and Appendix G details the SGIA conjunction.

6. Discussion and Conclusion

EditCLEVR illustrates how paired interventions complement segmentation and single-image decoding benchmarks. 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 object-attribute graph. SGIA rewards only rows where preservation, after-scene correctness, and the single-site change pattern all hold, whereas ΔSGIA and the component metrics help identify which part of the conjunction failed. The results suggest two related sources of error in these baselines. First, even with ground-truth instance masks, models still struggle under the CoGenT-derived OOD shift. Second, the choice of mask source affects native-model performance, but differences in masks alone do not explain the full performance gap.

Limitations and future work. EditCLEVR is synthetic and restricted to CLEVR-derived paired scenes, 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, generative, or VLM-based systems. The semantic protocol uses supervised probes. Appendix C repeats the semantic evaluation with an MLP probe to check sensitivity to probe capacity. This ablation changes the readout, while keeping the paired scenes, edit labels, and metric definitions fixed. Natural images, relational or continuous edits, simultaneous edits, stronger discovery sources, calibration baselines, and factor-structured readouts are natural next steps. The accompanying artifacts support use of EditCLEVR through a Blender-backed dataset generator, dataset download and evaluation tools, a reference ground-truth-mask encoder, and a 20k-pair dataset with instance masks, object attributes, edit metadata, difficulty tags, split metadata, and no-edit re-render controls.

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. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

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. URL https://arxiv.org/abs/2408.09162.

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. URL https://proceedings.mlr.press/v162/dittadi22a.html.

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*, 2020a. URL https://arxiv.org/abs/2012.05208.

Greff, K. et al. Multi-object datasets, 2020b. URL https://zenodo.org/records/4895643.

Greff, K. et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3761, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Greff_Kubric_A_Scalable_Dataset_Generator_CVPR_2022_paper.html.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019. URL https://doi.org/10.18653/v1/D19-1275.

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., Lawrence Zitnick, C., 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. URL https://doi.org/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. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

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. URL https://doi.org/10.1017/S0140525X16001837.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schoelkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

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. URL https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html.

Moon, Y.-B., Nam, H.-W., Choi, W., and Oh, T.-H. BEAF: Observing before-after changes to evaluate hallucination in vision-language models. In *Computer Vision – ECCV 2024*, pp. 232–248. Springer, 2024. doi: 10.1007/978-3-031-73247-8_14. URL https://doi.org/10.1007/978-3-031-73247-8_14.

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. In *Proceedings of the 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

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. URL https://arxiv.org/abs/2502.14786.

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). Each pair includes before/after RGB renders, before/after instance masks, before/after attribute records, edit metadata, and difficulty metadata. The hard-distractor split enforces a same-color and same-shape distractor, but it does not calibrate perceptual difficulty across examples; it should therefore be read as a controlled disambiguation stress test.

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; evaluation uses both predicted slot masks (for native rows and the hybrid) and slot features. SAM 2 (Ravi et al., 2025) 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 $\mathbb{R}^d$ (pooled patch tokens for ground-truth-mask rows; 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 the main paper; each predicted object can be assigned to at most one ground-truth instance within a frame. Semantic metrics use a $\mathrm{MatchBO} \geq 0.5$ gate on the edited object in both frames, while unconditioned variants keep the full denominator. The low-confidence rate in Table 5 is the fraction of rows excluded by this gate on ATOMIC_ID; it is small but nonzero (0.1–1.4% on ATOMIC_ID across native rows), so native semantic scores should be interpreted as faithfulness on matched objects rather than as a complete discovery-plus-faithfulness measure. 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 in Appendix C uses a 2-layer MLPClassifier (512×512, Adam, early stopping, seed 42) for all rows. 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 the edit suites for every model. The hard-distractor suite remains close to ATOMIC_ID for most rows, so it is best read as an instance-disambiguation diagnostic. The largest observed shift is the CoGenT OOD gap: every model loses strict SGIA on COGENT_OOD_CORE, and ΔSGIA drops for every model except SA. That exception is not evidence of better faithfulness: SA's strict SGIA is essentially zero on OOD-core, so the relaxed score is capturing partial single-site change patterns under weak after-scene correctness.

Table 2. Linear-probe SGIA / ΔSGIA across all edit 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 (GT-mask)	0.817	0.856	0.823	0.866	0.142	0.387	0.135	0.310	0.240
DINOv2 ViT-B/14 (GT-mask)	0.634	0.749	0.647	0.749	0.036	0.252	0.044	0.117	0.230
SigLIP 2 ViT-B/16 (GT-mask)	0.862	0.884	0.864	0.889	0.180	0.444	0.123	0.398	0.270
SA (conv, native)	0.011	0.125	0.016	0.144	0.001	0.091	0.000	0.163	0.073
DINOSAUR (native)	0.640	0.694	0.648	0.705	0.031	0.236	0.021	0.137	0.116
SAM 2 + DINO-S/8	0.485	0.560	0.488	0.571	0.136	0.273	0.111	0.237	0.264
SAM 2 + DINOv2	0.557	0.631	0.571	0.640	0.072	0.268	0.061	0.155	0.248
SAM 2 + SigLIP 2	0.512	0.583	0.508	0.580	0.093	0.289	0.061	0.228	0.284
DINOSAUR-mask + DINO-S/8	0.112	0.210	0.122	0.211	0.009	0.066	0.009	0.035	0.070

C. MLP Probe Replication

Table 3 reports the headline MLP values: SGIA, ΔSGIA, and TFA for every model. The stronger decoder generally raises ID scores but does not remove the OOD failure mode: OOD-core SGIA remains low for every row, and high OOD-core TFA can still coexist with weak full-scene intervention faithfulness. We therefore keep the linear probe as the primary protocol and use the MLP run as a probe-sensitivity check.

Table 3. Full MLP-probe headline results. 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 (GT-mask)	0.920	0.937	0.995	0.129	0.307	0.763
DINOv2 ViT-B/14 (GT-mask)	0.777	0.864	0.980	0.064	0.193	0.409
SigLIP 2 ViT-B/16 (GT-mask)	0.930	0.944	0.996	0.021	0.450	0.985
SA (conv, native)	0.269	0.363	0.903	0.000	0.237	0.872
DINOSAUR (native)	0.737	0.773	0.984	0.088	0.225	0.550
SAM 2 + DINO-S/8	0.519	0.591	0.940	0.073	0.187	0.716
SAM 2 + DINOv2	0.474	0.546	0.935	0.070	0.172	0.482
SAM 2 + SigLIP 2	0.529	0.602	0.937	0.035	0.228	0.939
DINOSAUR-mask + DINO-S/8	0.508	0.555	0.970	0.070	0.161	0.737

D. Soft-Mixture Ablation for SAM 2 Rows

The strict native protocol uses a one-to-one best-overlap assignment. As an ablation for the selected SAM 2 rows, we also replace this with an IoU-weighted convex combination of SAM 2 proposal features for each ground-truth object, $f_{\text{obj}} = \sum_k w_k s_k$ with normalized IoU weights w_k. Table 4 reports that soft mixtures improve ID scores for both SAM 2+DINO-S/8 and SAM 2+SigLIP 2, and they also improve OOD-core ΔSGIA. The gains are not enough to remove the CoGenT gap, so the strict protocol remains the headline setting and the soft mixture is treated as a proposal-aggregation ablation.

Table 4. Strict vs. soft IoU-mixture proposal aggregation for selected SAM 2 rows (linear probes).

	ATOMIC_ID		COGENT_OOD_CORE			
Model	SGIA	ΔSGIA	SGIA	ΔSGIA	EOA(ID)	NED ↓
SAM 2 + DINO-S/8 (strict)	0.485	0.560	0.111	0.237	0.713	0.264
SAM 2 + DINO-S/8 (soft)	0.635	0.700	0.132	0.292	0.810	0.243
SAM 2 + SigLIP 2 (strict)	0.512	0.583	0.061	0.228	0.701	0.284
SAM 2 + SigLIP 2 (soft)	0.678	0.734	0.073	0.301	0.769	0.270

E. Native Discovery Diagnostics

Table 5 reports the discovery-side diagnostics that gate the semantic metrics for native and hybrid rows: foreground ARI (FG-ARI), best-overlap coverage (MatchBO), matched-mask IoU (MatchIoU), and low-confidence rate. The low-confidence rate is the fraction of rows excluded by the edited-object MatchBO gate on ATOMIC_ID. DINOSAUR masks have very high best-overlap but low IoU, meaning they reliably include the correct object while covering extra pixels. SAM 2 masks have lower best-overlap than DINOSAUR masks but substantially higher IoU, which is consistent with their stronger frozen-feature results in the main-paper table. SA has the highest IoU in this table but a much larger low-confidence rate, so its semantic failures cannot be reduced to mask size alone.

F. Per-Factor and Object-Count Linear Breakdowns

Table 6 breaks the linear-probe results down by edited factor and by scene object count, pooling the three non-derived edit suites (ATOMIC_ID, HARD_DISTRACTOR, and COGENT_OOD). Each cell reports SGIA/ΔSGIA. The object-count columns show lower scores in larger scenes: strict SGIA falls from 0.738 to 0.601 for the SigLIP 2 GT-mask, from 0.589 to 0.295 for

Table 5. Discovery diagnostics on ATOMIC_ID. Higher is better except low-confidence rate. MatchBO is the best-overlap coverage score used for gates; MatchIoU is the IoU of the matched pair. Low-confidence is the MatchBO-gate exclusion rate for semantic metrics.

Model	FG-ARI	MatchBO	MatchIoU	LowConf ↓
SA (conv)	0.869	0.912	0.420	0.014
DINOSAUR	0.975	0.983	0.098	0.002
SAM 2 + DINO/DINOv2/SigLIP 2	0.913	0.974	0.270	0.001
DINOSAUR-mask + DINO-S/8	0.975	0.983	0.098	0.002

DINOSAUR native, and from 0.527 to 0.204 for SAM 2+SigLIP 2. The factor columns show that failures are not isolated to a single factor; material and size are sometimes easier than color/shape for individual rows, but the OOD and object-count effects remain visible across the table.

Table 6. Per-factor and per-object-count breakdowns for linear probes. Each cell is SGIA/ΔSGIA, pooled over ATOMIC_ID, HARD_DISTRACTOR, and COGENT_OOD; NO_EDIT and the derived OOD-core slice are excluded. These pooled values are descriptive because they mix ID, hard-distractor, and full CoGenT-OOD suites.

	Edited factor			
Model	Color	Material	Size	Shape
DINO ViT-S/8 (GT-mask)	0.627/0.712	0.628/0.759	0.616/0.704	0.633/0.724
DINOv2 ViT-B/14 (GT-mask)	0.491/0.613	0.464/0.652	0.466/0.603	0.447/0.560
SigLIP 2 ViT-B/16 (GT-mask)	0.682/0.770	0.658/0.764	0.646/0.726	0.684/0.779
SA (conv, native)	0.016/0.167	0.008/0.084	0.011/0.174	0.004/0.059
DINOSAUR (native)	0.471/0.545	0.470/0.604	0.471/0.567	0.463/0.548
SAM 2 + DINO-S/8	0.395/0.477	0.400/0.514	0.362/0.446	0.388/0.488
SAM 2 + DINOv2	0.442/0.535	0.430/0.572	0.417/0.517	0.402/0.495
SAM 2 + SigLIP 2	0.414/0.513	0.400/0.520	0.359/0.460	0.393/0.499
DINOSAUR-mask + DINO-S/8	0.095/0.182	0.082/0.174	0.085/0.155	0.079/0.165
	Objects in scene			
Model	3	4	5	6
DINO ViT-S/8 (GT-mask)	0.700/0.786	0.633/0.729	0.617/0.717	0.554/0.666
DINOv2 ViT-B/14 (GT-mask)	0.561/0.673	0.495/0.631	0.452/0.592	0.361/0.532
SigLIP 2 ViT-B/16 (GT-mask)	0.738/0.828	0.691/0.770	0.641/0.732	0.601/0.710
SA (conv, native)	0.030/0.229	0.008/0.159	0.001/0.075	0.001/0.022
DINOSAUR (native)	0.589/0.677	0.538/0.645	0.451/0.550	0.295/0.392
SAM 2 + DINO-S/8	0.525/0.599	0.455/0.543	0.371/0.474	0.194/0.309
SAM 2 + DINOv2	0.584/0.675	0.522/0.625	0.385/0.497	0.199/0.323
SAM 2 + SigLIP 2	0.527/0.622	0.461/0.570	0.373/0.471	0.204/0.330
DINOSAUR-mask + DINO-S/8	0.144/0.251	0.095/0.178	0.076/0.156	0.027/0.092

G. What Gets Counted as a Failure

SGIA is intentionally strict: a row passes only when the after-frame object-attribute 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. Component terms can look strong even when the joint event fails: SigLIP 2 GT-mask TFA is 0.995 on ID and 0.980 on OOD-core, yet OOD-core SGIA is only 0.123; SA has OOD-core TFA 0.863 and ΔSGIA 0.163, yet SGIA is 0.000. We therefore use SGIA as the headline strict measure, while using ΔSGIA, TFA, NFP, UOP, the MLP check, and discovery diagnostics to separate decoding noise, intervention-pattern errors, probe effects, and object-assignment failures.

H. Reproducibility

The accompanying code and dataset artifacts are available at https://github.com/torux-bughunter/EditCLEVR. The package includes Python tools for the Blender-backed CLEVR generator, dataset download, metric evaluation, the standardized `Evaluator`, and the reference ground-truth-mask encoder. The released dataset contains 20,000 paired examples, including no-edit re-render controls, with before/after RGB images, instance masks, scene JSON, object attributes, edit metadata, difficulty tags, and `splits.json`. The evaluation pipeline trains attribute probes on the train split and reports headline metrics with bootstrap confidence intervals, `results.json`, and per-split/suite CSVs.