Object-centirc Vision Language Model

Weizhen Zhou[1*] **Yuheng Wu**[1*] **Xingyu Han**[1*]

[1] New York University
wz3008@nyu.edu, yw5372@nyu.edu, xh2787@nyu.edu

Abstract

Our work is inspired by a recent line of research that aims to train AI models from a small amount of input from a single child, rather than astronomical amounts of data from the web. While such approaches have demonstrated promising results in multimodal understanding, they remain poor at segmenting the visual world into objects and at using object-based representations for learning, reasoning, and prediction. In this work, we explore whether incorporating rich, compositional object representations into the model architecture can lead to better data efficiency and more human-like cognitive development. Furthermore, we investigate a more fundamental question: whether object-centric representations are inherently more data-efficient than non-slot visual tokens (e.g., ViT tokens) for multimodal understanding. The related code is available at https://github.com/zhouwzh/OCL_VLM.

1 Introduction

Modern AI systems learn in ways that are fundamentally different from how humans acquire knowledge(Frank 2023). State-of-the-art models are typically trained on massive-scale datasets with word counts in the billions or trillions, whereas human children are able to develop comparable levels of understanding from extremely limited linguistic exposure. This discrepancy is often referred to as the data gap between artificial and human learning.

Motivated by this gap, a recent line of work has emerged under the name of Single-Child Development AI (SCAI) models(Wang et al. 2023; Vong et al. 2024; Orhan and Lake 2024), which aim to train AI systems from the first-person visual perspective of a single young child. These models are commonly trained on proxy egocentric data such as the SAY-Cam dataset, which contains long-term head-mounted camera recordings of children's daily visual experience. While SCAI models have demonstrated promising results in representation learning and multimodal grounding, they still face significant challenges, particularly in forming robust and disentangled object-level representations in complex, natural scenes.

In parallel, another line of research—Object-Centric Learning (OCL)(Wu, Lee, and Ahn 2024)—has shown strong advantages in discovering and representing individual objects in an unsupervised manner. By introducing structured inductive biases such as slot-based representations, OCL methods have achieved substantial improvements in object discovery, compositionality, and generalization, especially on synthetic or controlled video datasets.

In this project, we aim to explore whether object-centric learning can help address the object representation weaknesses observed in SCAI-style models when applied to natural, egocentric video data. Furthermore, we investigate a more fundamental question: whether structured, object-level representations can serve as a more effective interface for multimodal learning compared to holistic scene-level representations.

Our main contributions are as follows:

- Reproduction and evaluation of prior work. We reproduce the Grounded language acquisition through the eyes and ears of a single child(CVCL) paper(Vong et al. 2024) as well as several representative OCL methods, including STEVE(Singh, Wu, and Ahn 2022), SlotContrast(Manasyan et al. 2025) and MetaSlot(Liu et al. 2025). We evaluate these models using linear probing, four-way classification (distinguishing object from given label), and *Foreground Adjusted Rand Index* (FG-ARI) for object discovery quality.

- Object-centric multimodal learning. We design a CLIP-style multimodal training framework that combines object-centric visual representations with language, enabling slot-based vision–language contrastive learning.

- Test-time training for OCL models. We conduct test-time training experiments on selected OCL methods and analyze their effects on linear probing and four-way classification performance.

- Object-centric representations for vision–language pairing. We construct object-centric representation–language pairs from the MOVi-A dataset and compare slot-based OCL representations with ViT token-based representations in multimodal learning settings. Through controlled experiments, we analyze their differences in representation quality and downstream performance.

[*]These authors contributed equally.

2 Related Work

Single-Child Egocentric Learning Models

Recent work has explored learning visual and multimodal representations from the perspective of a single child, aiming to bridge the gap between large-scale web-trained models and human-like learning. The CVCL framework proposes to learn visual representations from egocentric video streams using self-supervised objectives, demonstrating that meaningful representations can emerge from limited, child-centered experience. However, these models primarily rely on holistic scene-level representations and often lack explicit mechanisms for disentangling individual objects, which limits their ability to reason about object-level structure and compositionality in complex natural scenes.

Object-Centric Learning (OCL)

Object-centric learning has emerged as a powerful paradigm for unsupervised object discovery, typically by representing scenes as a set of slots that correspond to individual objects. STEVE(Singh, Wu, and Ahn 2022) extends slot-based learning to complex and naturalistic videos, showing that temporal consistency can significantly improve object discovery. MetaSlot(Liu et al. 2025) further addresses the limitation of a fixed number of slots by dynamically adapting slot allocation to scene complexity. SlotContrast(Manasyan et al. 2025) introduces contrastive objectives at the object level, encouraging discriminative and consistent slot representations across views and time. Despite their success on synthetic datasets and controlled settings, these methods often struggle to generalize to natural egocentric videos with cluttered backgrounds and severe viewpoint changes.

Object-Centric and Vision–Language Models

More recent work has investigated incorporating object-centric representations into vision–language models (VLMs). CTRL-O(Didolkar et al. 2025) proposes language-controllable object-centric visual representation learning by introducing text supervision into the slot learning process, combined with a contrastive learning stage to align slots with linguistic concepts. This approach demonstrates improved controllability and interpretability of object-centric representations in multimodal settings. OC-VTP(Li et al. 2025) focuses on improving the efficiency of VLMs by introducing a plug-and-play module for object-centric vision token pruning, selectively retaining object-relevant ViT tokens while discarding redundant background information. Concept-Guided Self-Supervised Learning (CG-SSL)(Atito et al.) further explores using high-level semantic concepts to guide representation learning without explicit labels, although details of the implementation and empirical analysis remain limited due to the lack of publicly available code.

3 Paper Reproduction

3.1 CVCL Reproduction.

We reproduced the CVCL training pipeline on the SAYCam-S dataset from Grounded language acquisition through the eyes and ears of a single child. CVCL learns aligned visual–linguistic embeddings from temporally paired child-directed utterances and egocentric video frames using a symmetric contrastive objective. Concretely, we used a ResNeXt-50 32x4d vision encoder with a 512-d projection head, where the backbone was initialized from DINO pretraining on the same child's egocentric videos and kept frozen during CVCL training. The language encoder was implemented as a 512-d word embedding table, and utterance embeddings were obtained by averaging token embeddings. For optimization, we trained up to 400 epochs with AdamW (lr = 1e-4, weight decay = 0.1), batch size 8, temperature $\tau = 0.07$, ReduceLROnPlateau scheduling (factor 0.1, patience 20), and early stopping based on validation contrastive loss. We evaluated the reproduced model on the four-way classification protocol (Labeled-S), achieving **62.9%** accuracy, consistent with the result in the original paper.

Figure 1: MetaSlot visualization.

3.2 OCL Paper Reproduction

We also reproduce the result of some OCL paper, e.g. STEVE, Slotcontrast, Metaslot, respectively trian them on their dataset, and later apply on SAYCam-S dataset.

STEVE: We first reproduced STEVE on the synthetic video benchmark MOVi-E dataset. STEVE employs a recurrent slot encoder for object-centric video representation learning: for each frame, slots are refined with Slot Attention and updated via a GRU (2 refinement iterations per frame), followed by a 1-layer Transformer with 4 attention heads to model slot interactions. On the reconstruction side, STEVE

uses a discrete VAE (dVAE) to tokenize images into a sequence of discrete codes using 4×4 patches with a vocabulary size of 4096, and trains an autoregressive Transformer decoder for reconstruction. We adopted the hyperparameters reported in the appendix: batch size $= 24$, episode length $= 3$, and trained for 200K steps with slot size $= 192$. The learning rate schedule uses a 30K-step warmup followed by exponential decay (half-life 250K steps), and the dVAE Gumbel-Softmax temperature is annealed from 1.0 to 0.1 over the first 30K steps. With this setup, we reproduced the reported *Video FG-ARI* on MOVi-E, achieving 53.371% as same in the original paper.

Slotcontrast: We reproduced SlotContrast on the MOVi-C dataset.SlotContrast builds upon a recurrent slot-based video object-centric architecture and introduces a temporal slot–slot contrastive loss to explicitly enforce temporal consistency. Specifically, each video frame is first encoded using a frozen DINOv2 ViT backbone, whose dense patch features are adapted via a lightweight MLP before being grouped into slots by a recurrent Slot Attention module. Slots from the previous frame are used to initialize the current frame, and a learned global initialization is adopted for the first frame. In addition to the standard feature reconstruction loss in self-supervised feature space, SlotContrast introduces a contrastive InfoNCE objective that attracts corresponding slots across consecutive frames while repelling all other slots within the batch. The final training objective is a weighted combination of the reconstruction loss and the slot–slot contrastive loss. Following the appendix, we used DINOv2 ViT-B/14 for MOVi-E, slot dimension $= 128$, input resolution 336×336, and trained the model on full video sequences of 24 frames. With this setup, we reproduced the reported object discovery performance on MOVi-C, achieving a *Video FG-ARI* of **73.1%**, consistent with the results reported in the original paper.

MetaSlot: We further reproduced MetaSlot on COCO dataset. MetaSlot is a plug-and-play variant of Slot Attention that enables dynamic slot allocation through a prototype-based vector-quantized codebook. The method employs a two-stage aggregation process: in the first stage, intermediate slots are produced via Slot Attention and matched to a global prototype codebook using nearest-neighbor quantization, after which duplicate slots are pruned. In the second stage, the remaining prototype slots are refined using masked Slot Attention with progressively annealed Gaussian noise, which stabilizes optimization and improves object binding. The prototype codebook is updated using a mini-batch K-means strategy with exponential moving average, while gradients are truncated across stages to ensure stable training. Following the appendix, all models were trained using a DINOv2 ViT backbone with shared data augmentation and hyperparameters, trained for 50K steps with Adam optimizer, batch size $= 32$, and a fixed codebook size of 512. Using this setup, we successfully reproduced the reported improvements of MetaSlot over standard Slot Attention on object discovery benchmarks, achieving *FG-ARI* of **41.2%** on COCO dataset, consistent with the trends reported in the original work. A sample visualization of the slot of MetaSlot in shown in Figure1.

Method	Accuracy (%)
ResNext50	88.5
ViTb16	82.6
Steve Encoder	72.9
Slotcontrast Encoder	74.07
MetaSlot Encoder	77.05

Table 1: Linear Probe Results

Method	TopK (%)	SoftAlign (%)
CVCL-Steve	54.5	49.6
CVCL-Slotcontrast	58.23	49.95
CVCL-MetaSlot	**64.8**	58.86

Table 2: Results of four-way classification, with the baseline of CVCL has a result of **62.9%**.

4 OCL with multimodal learning

In this section, we describe how we apply object-centric learning (OCL) models to multimodal learning from video–utterance pairs based on the SAYCam dataset.

4.1 Linear Probe across Different Backbones

Since different object-centric models adopt different visual backbones—including ViT-B/14, ViT-B/16, and ResNeXt-50—a direct comparison of downstream performance can be confounded by backbone capacity. To control for this factor, we conducted linear probing experiments on SAYCam-pretrained visual backbones and on the object-centric representations produced by their corresponding OCL models. For each model, we freeze the backbone (or slot encoder) and train a linear classifier on top of the extracted representations using a **cross-entropy loss**. When probing slot-based models, we aggregate slot features by taking average of them to obtain a fixed-dimensional representation. All linear probes are trained for **200** epochs using the Adam optimizer with learning rate **1e-3**, batch size **128**, and early stopping based on validation accuracy. As shown in Table 1, slot-based representations exhibit slightly lower accuracy compared to their corresponding backbone features; however, the performance gap remains modest. Notably, object-centric slot representations are competitive with strong pre-trained visual backbones, suggesting that the learned slots preserve most of the discriminative information while providing structured object-level representations. We hypothesize that the observed performance drop is primarily due to the slot selection and aggregation process, which prioritizes object-centric decomposition over global scene information.

4.2 Multimodal Evaluation with OCL Encoders

To assess the role of object-centric representations in multimodal learning, we replaced the vision encoder in CVCL with the vision encoders from different OCL models. In particular, when using a Slot-Attention-based video encoder (SAVi), the model outputs slot features $\{s_i\}_{i=1}^{K}$ and the corresponding pixel-level attention masks $\{m_i\}_{i=1}^{K}$ for each frame. Since the slot set includes both foreground and back-

ground components, we explored multiple strategies to select or aggregate the most relevant slots for vision–language alignment, including **TopK** and **Soft-Alignment**. Quantitative results are summarized in Table 2.

TopK Slot Selection. After a warm-up stage, we compute the similarity between each slot feature and the paired text feature, and select the top-k most relevant slots. Formally, let the slot-conditioned image features be $I \in R^{B\times K\times D}$ and text features be $T \in R^{B\times D}$. For each sample i and slot j, we compute a similarity score

$$s_{ij} = \langle I_{i,j,:}, T_{i,:} \rangle \in R, \tag{1}$$

yielding $\mathbf{s} \in R^{B\times K}$. We then select the top-k slots per sample,

$$\mathcal{J}_i = \mathrm{TopK}(\mathbf{s}_{i,:}), \qquad I_i^{\mathrm{sel}} = \{I_{i,j,:} \mid j \in \mathcal{J}_i\} \in R^{k\times D}, \tag{2}$$

and apply the same InfoNCE loss as in the warm-up stage using the selected slots.

Soft-Alignment over Slots. Instead of hard selection, we compute a soft alignment distribution over slots and form a weighted sum of slot features, which is then used in a CLIP-style symmetric InfoNCE loss. Using the same similarity scores s_{ij}, we define attention weights

$$w_{ij} = \frac{\exp(s_{ij})}{\sum_{m=1}^{K} \exp(s_{im})} \in R, \qquad \mathbf{w} \in R^{B\times K}, \tag{3}$$

and aggregate slot features as

$$V_i = \sum_{j=1}^{K} w_{ij}\, I_{i,j,:} \in R^D, \qquad V \in R^{B\times D}. \tag{4}$$

We then compute CLIP-style logits

$$\mathrm{logits}_{\mathrm{img}} = \frac{1}{\tau} V T^\top \in R^{B\times B}, \tag{5}$$

$$\mathrm{logits}_{\mathrm{text}} = \frac{1}{\tau} T V^\top \in R^{B\times B}, \tag{6}$$

and minimize the symmetric InfoNCE objective

$$L_{\mathrm{img}} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp(\mathrm{logits}_{\mathrm{img}}^{i,i})}{\sum_{j=1}^{B}\exp(\mathrm{logits}_{\mathrm{img}}^{i,j})}, \tag{7}$$

$$L_{\mathrm{text}} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp(\mathrm{logits}_{\mathrm{text}}^{i,i})}{\sum_{j=1}^{B}\exp(\mathrm{logits}_{\mathrm{text}}^{i,j})}, \tag{8}$$

$$L = \frac{1}{2}\left(L_{\mathrm{img}} + L_{\mathrm{text}}\right). \tag{9}$$

4.3 FG-ARI Evaluation

Since the SAYCam dataset does not provide ground-truth instance segmentation masks, we adopt *Segment Anything* (SAM) to generate pseudo ground-truth masks for evaluation. Specifically, we apply SAM to each frame to obtain a set of segmentation masks. An example of the resulting pseudo masks is illustrated in Figure 2. Using these SAM-generated masks as a proxy for ground truth, we evaluate the object discovery quality of slot masks produced by STEVE and MetaSlot using the FG-ARI. Quantitative results are reported in Table 3, showing that object-centric models exhibit noticeably lower FG-ARI on SAYCam compared to their performance on synthetic datasets such as MOVi-E, indicating that OCL methods remain substantially more challenging to apply to natural, egocentric videos with complex backgrounds and viewpoint variations.

Number of Slots	Steve	Metaslot
6	43.025	/
8	42.109	39.69
11	42.265	/
Untrained Model	16.527	9.46

Table 3: Result of FG-ARI evaluation on SAYCam with using the mask from SAM as groundtruch.

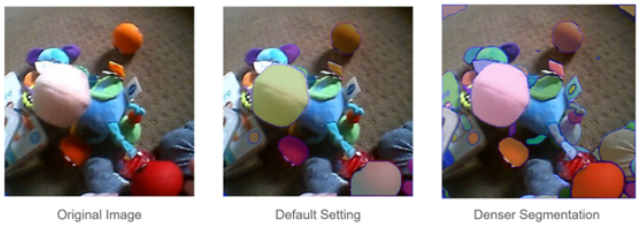

Figure 2: SAM Generated SAYCam masks.

5 OCL multimodal with Test Time Traing

As discussed above, on natural datasets such as SAYCam, existing OCL methods do not consistently outperform backbone representations in linear probing or four-way classification. In some cases, object-centric models achieve marginal improvements, but the overall gains remain limited. Moreover, the FG-ARI results indicate that object discovery quality on natural, egocentric videos is still slightly lower than on synthetic datasets. These observations suggest that models trained on synthetic or curated data distributions may not generalize optimally to the complex visual statistics of natural videos.

Motivated by this gap, we adopt *test-time training* (TTT), a technique that adapts model parameters at inference time using self-supervised objectives. TTT has been shown to improve robustness under distribution shift by allowing the model to dynamically adjust to test data without requiring additional annotations.

To evaluate whether TTT can improve FG-ARI performance, we applied TTT to a small subset of the Labeled-S dataset using the MetaSlot model, and conducted a hyperparameter search over the learning rate $\{2\times10^{-6}, 4\times10^{-6}, 6\times10^{-6}, 8\times10^{-6}, 1\times10^{-5}, 2\times10^{-5}, 6\times10^{-5}, 8\times10^{-5}, 1\times10^{-4}, 1.2\times10^{-4}\}$ and the number of update steps $\{1, 2, 4, 6, 8, 10, 12, 14, 16, 18\}$. The result is shown in Figure3. Furhter, we tried TTT on inference stage of a CVCL-Metaslot trained on SAYCAM. With learning rate = 2×10^{-5}, step = 6. Since CVCL doesn't have a image-only unsupervised loss, so during TTT stage we use the unsu-

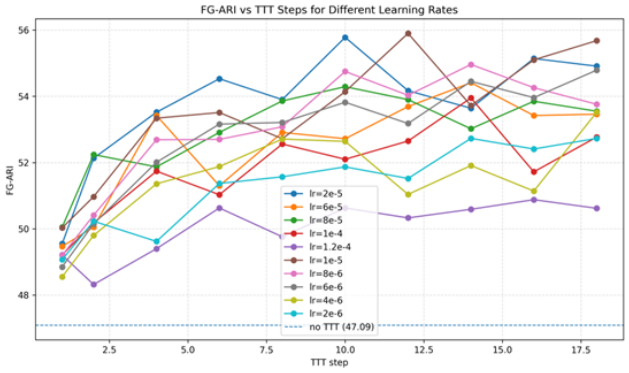

Figure 3: FG-ARI vs TTT Steps for Different lr

	Accuracy(%)
CVCL-MetaSlot w.o. TTT	64.8
CVCL-MetaSlot w. TTT	64.86

Table 4: TTT on CVCL-Slot contrastive learning - four way classification Evaluation.

pervised feature reconstruction loss of MetaSlot model. As shown in Table:4, the improvment is limited.

We observe that TTT yields large performance improvements in FG-ARI; however, the overall gains remain limited. We attribute this behavior to several intrinsic challenges of the SAYCam dataset

- The background textures in the SAYCam dataset images are notably richer.
- The background in the SYACam dataset is not remain still due to the ego-centric view.
- SYACAM dataset use a much lower fps, objects moves faster
- SAYCam dataset's objects are not always locate in central of the picture.

6 Slot representation v.s. ViT tokens

Therefore, we want to take a step back and study whether slot-based representations are fundamentally more data-efficient than non-slot visual tokens(e.g., ViT tokens) in terms of aligning the text and the objects.

To achieve this goal, we select MOVi-A as a simplified dataset for experiments. Each frame in MOVi-A is annotated with bounding boxes, object-level textual descriptions, and ground-truth segmentation masks. Leveraging these annotations together with the slot representations produced by STEVE, we construct a new dataset that pairs slot features with object captions. Concretely, for each frame, we associate each ground-truth object caption with the corresponding slot representation inferred by the OCL model. Each sample in the resulting dataset consists of the set of object captions present in a frame, the slot representation corresponding to each object, the groundtruth mask and the mask provided by OCL model. An example of the constructed slot–caption pairs is illustrated in Figure 4.

	Accuracy(%)
ViTb16	53.31
STEVE	**63.16**

Table 5: Zero-shot text-to-object retrieval performance of slot representations and ViT features.

Building on the constructed slot–caption, we adopt a CLIP-style multimodal contrastive learning framework to align visual representations with text embeddings. Specifically, we directly contrast slot features or ViT token features against their corresponding textual descriptions using the InfoNCE loss. The objective is to encourage object-centric slot representations to better capture object-level semantics compared to token-level representations. We evaluate the learned representations under zero-shot settings, including text-to-object retrieval, and compare the performance of slot representations and ViT features. As shown in Table 5, slot-based representations demonstrate stronger generalization ability under limited training data regimes, outperforming ViT token representations in zero-shot evaluation.

7 Conclusion

Our project aims not only to practice deep learning techniques within the scope of existing knowledge, but also to leverage them to explore a problem that remains at the research frontier of the broader academic community: how to learn robust, object-centric representations from natural, egocentric visual data and integrate them effectively into multimodal understanding.

Due to the inherent limitations of current object-centric learning (OCL) methods, their ability to handle highly naturalistic datasets such as SAYCam remains constrained, leading to suboptimal object segmentation and discovery performance. While our experiments demonstrate modest improvements on certain downstream tasks, these gains are not sufficient to close the gap between synthetic and real-world settings. Nevertheless, the observed improvements suggest that object-centric representations retain meaningful semantic structure even under challenging data distributions.

At the same time, by starting from a simplified and well-controlled setting, we validate a fundamental hypothesis: explicit object-centric representations can facilitate stronger alignment between visual features and object-level language descriptions compared to token-based representations. This finding highlights the potential of slot-based representations as a more structured and interpretable interface between vision and language.

Therefore we believe that further progress will require combining object-centric inductive biases with more scalable training strategies, stronger temporal modeling, and adaptive mechanisms such as test-time training. Such directions may ultimately enable OCL methods to better bridge the gap between synthetic benchmarks and complex natural environments, advancing their applicability to real-world multimodal learning scenarios.

References

Atito, S.; Kittler, J.; Razzak, I.; and Awais, M. ???? CG-SSL: Concept-Guided Self-Supervised Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Didolkar, A.; Zadaianchuk, A.; Awal, R.; Seitzer, M.; Gavves, E.; and Agrawal, A. 2025. Ctrl-o: language-controllable object-centric visual representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29523–29533.

Frank, M. C. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11): 990–992.

Li, G.; Zhao, R.; Deng, J.; Wang, Y.; and Pajarinen, J. 2025. Object-Centric Vision Token Pruning for Vision Language Models. *arXiv preprint arXiv:2511.20439*.

Liu, H.; Zhao, R.; Chen, H.; and Pajarinen, J. 2025. MetaSlot: Break Through the Fixed Number of Slots in Object-Centric Learning. *arXiv preprint arXiv:2505.20772*.

Manasyan, A.; Seitzer, M.; Radovic, F.; Martius, G.; and Zadaianchuk, A. 2025. Temporally consistent object-centric learning by contrasting slots. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5401–5411.

Orhan, A. E.; and Lake, B. M. 2024. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, 6(3): 271–283.

Singh, G.; Wu, Y.-F.; and Ahn, S. 2022. Simple unsupervised object-centric learning for complex and naturalistic videos. *Advances in Neural Information Processing Systems*, 35: 18181–18196.

Vong, W. K.; Wang, W.; Orhan, A. E.; and Lake, B. M. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682): 504–511.

Wang, W.; Vong, W. K.; Kim, N.; and Lake, B. M. 2023. Finding structure in one child's linguistic experience. *Cognitive science*, 47(6): e13305.

Wu, Y.-F.; Lee, M.; and Ahn, S. 2024. Structured world modeling via semantic vector quantization. *CoRR*.

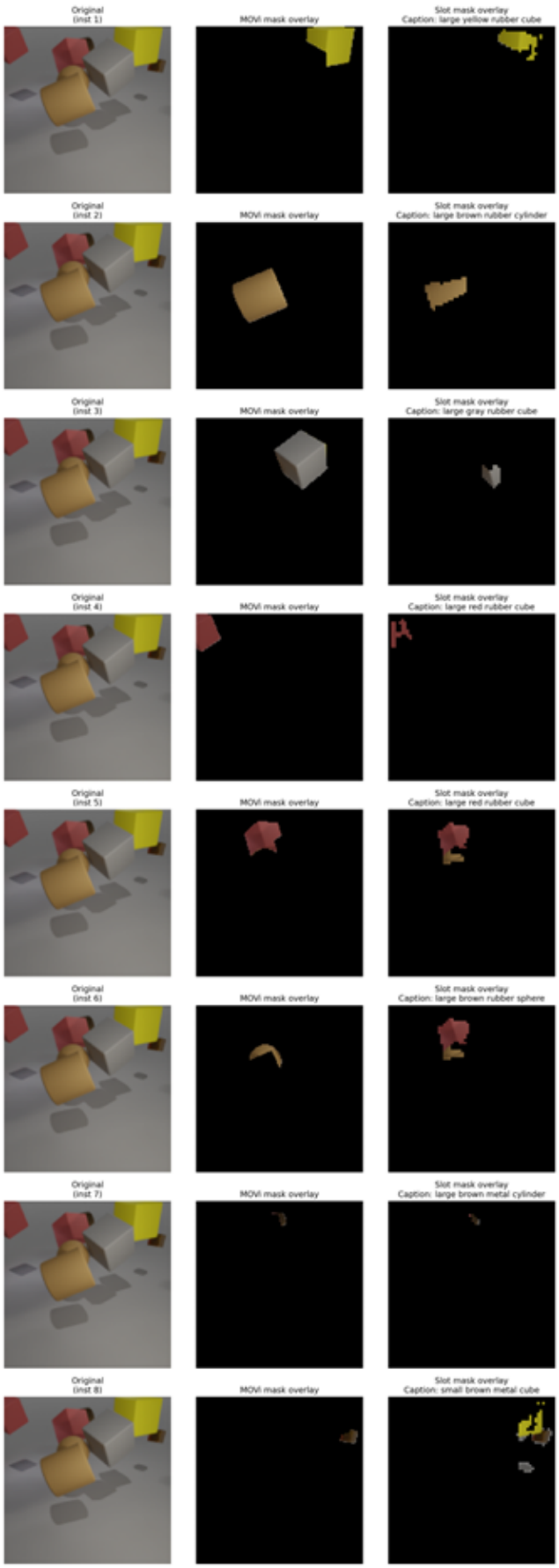

Figure 4: MOVI-A paired data sample.