Why Can't I Open My Drawer? Mitigating Object-Driven Shortcuts in Zero-Shot Compositional Action Recognition

Geo Ahn[1]*, Inwoong Lee[2], Taeoh Kim[2], Minho Shim[2], Dongyoon Wee[2], and Jinwoo Choi[1]†

[1]Kyung Hee University [2]NAVER Cloud
{ahngeo11, jinwoochoi}@khu.ac.kr
{inwoong.lee, taeoh.kim, minho.shim, dongyoon.wee}@navercorp.com

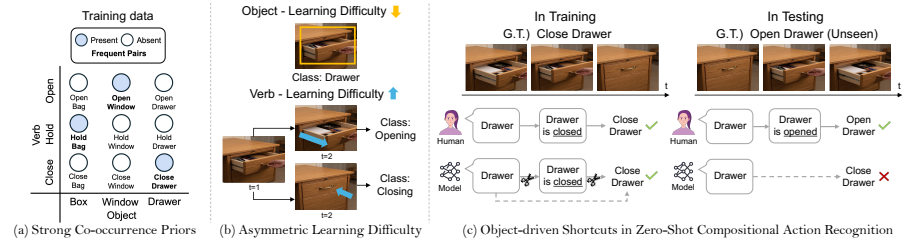
Abstract. Zero-Shot Compositional Action Recognition (ZS-CAR) requires recognizing novel verb–object combinations composed of previously observed primitives. In this work, we tackle a key failure mode: models predict verbs via *object-driven shortcuts* (*i.e.*, relying on the labeled object class) rather than temporal evidence. We argue that sparse compositional supervision and verb–object learning asymmetry can promote object-driven shortcut learning. Our analysis with proposed diagnostic metrics shows that existing methods overfit to training co-occurrence patterns and underuse temporal verb cues, resulting in weak generalization to unseen compositions. To address object-driven shortcuts, we propose Robust COmpositional REpresentations (`RCORE`) with two components. Co-occurrence Prior Regularization (CPR) adds explicit supervision for unseen compositions and regularizes the model against frequent co-occurrence priors by treating them as hard negatives. Temporal Order Regularization for Composition (TORC) enforces temporal-order sensitivity to learn temporally grounded verb representations. Across Sth-com and EK100-com, `RCORE` reduces shortcut diagnostics and consequently improves compositional generalization. The code is available at https://github.com/KHU-VLL/RCORE.

Keywords: Zero-shot compositional action recognition · Compositional generalization · Video representation learning

1 Introduction

Many human actions can be decomposed into two semantic primitives—verbs and objects—and robust video understanding requires recognizing each component and reasoning over their interaction as a composition [2, 7, 8, 12, 27]. Zero-Shot Compositional Action Recognition (ZS-CAR) formalizes this goal: a model must recognize unseen verb–object pairs while keeping the verb and object vocabularies shared across splits [15, 16, 19, 43].

*This work was done during an internship at NAVER Cloud.
†Corresponding author.



(a) Strong Co-occurrence Priors (b) Asymmetric Learning Difficulty (c) Object-driven Shortcuts in Zero-Shot Compositional Action Recognition

Fig. 1: Why object-driven shortcuts emerge in compositional video understanding? (a) **Co-occurrence bias.** Datasets are intrinsically sparse and highly skewed in their verb–object combinations, creating strong co-occurrence priors. (b) **Asymmetric learning difficulty.** Objects are visually explicit and easy to recognize from a single frame, whereas verbs require multi-frame temporal reasoning. This imbalance makes object features dominate training signals. (c) **Object-driven shortcuts.** Together, these two factors drive models to adopt *object-driven shortcuts*, hindering genuine compositional generalization. Once the object is recognized, the model often predicts the most frequent verb paired with it, ignoring temporal evidence.

In this setting, a key failure mode emerges: models often predict verbs via *object-driven shortcuts* rather than temporal evidence. Throughout, *object* refers to the *labeled noun class* in the verb–object vocabulary (not unlabeled background entities). We argue that object-driven shortcuts arise from two issues inherent to ZS-CAR. First, **compositional supervision is sparse and skewed**. As illustrated in Figure 1 (a), datasets cover only a small fraction of the combinatorial verb–object space, and training labels concentrate on a limited set of frequent pairs [7,27]. This skew induces a strong *co-occurrence prior* in the *training data*—the empirical dominance of frequent verb–object pairs—which can be amplified into a *model shortcut* at training time. Under such skewed supervision, predictions for unseen compositions tend to drift toward *seen* (especially frequent) pairs. Second, **verbs are harder to learn than objects**. As shown in Figure 1 (b), an object often becomes recognizable from a single frame, whereas a verb typically requires multi-frame temporal reasoning. Prior work on shortcut learning suggests that models tend to rely on easier-to-learn cues rather than the intended evidence [11, 20, 30, 35, 37]. In ZS-CAR, this verb–object difficulty gap makes object cues an especially attractive shortcut for verb prediction.

In this work, we posit that these two intertwined issues in verb-object compositional learning promote *object-driven shortcuts*, which we diagnose as the key factor hindering ZS-CAR. As Figure 1 (c) illustrates, a model can match a verb once it recognizes its frequently co-occurring object during training, without modeling the temporal dynamics essential to correctly recognize the verb. This behavior weakens generalization to unseen compositions because it ties verb prediction to object-dependent priors rather than temporally grounded evidence. We observe that models [19], even with a modern video-pretrained backbone [40], indeed confuse verbs with opposite temporal order (*e.g.*, opening *vs*. closing) on unseen compositions (see Section 5.2).

To dissect how object-driven shortcuts hinder unseen generalization, we perform a comprehensive quantitative diagnosis. First, through controlled experi-

ments, we show that the asymmetric learning difficulty between verbs and objects, together with skewed co-occurrence priors, promotes verb shortcut learning driven by object cues. Next, we *quantify* shortcut-driven failures in existing ZS-CAR models by two diagnostic ratios: False Seen Prediction (FSP) and False Co-occurrence Prediction (FCP), which measure how often an unseen input collapses to a seen—and especially frequent—training composition. Using FSP/FCP, we show that a strong ZS-CAR baseline [19] remains heavily biased toward seen compositions, resulting in poor generalization to unseen compositions.

Driven by the diagnosis, we propose Robust COmpositional REpresentations (`RCORE`), targeting the two root causes of object-driven shortcuts. First, Co-occurrence Prior Regularization (CPR) expands supervision over originally absent compositions and suppresses the dominance of frequently seen pairs by treating them as hard negatives. Second, Temporal Order Regularization for Composition (TORC) enforces temporal-order sensitivity so that the model learns temporally grounded verb representations instead of relying on static cues.

We validate `RCORE` on Sth-com [19] under a realistic open-world evaluation protocol that evaluates all possible verb-object pairs, without requiring test-set ground truth labels. We also introduce EK100-com, a ZS-CAR dataset repurposed from EPIC-KITCHENS-100 [7], which exhibits more severe compositional sparsity than Sth-com [19]. Across datasets and backbones, `RCORE` reduces shortcut diagnostics (FSP/FCP) and improves unseen compositional generalization.

In this work, we make the following major contributions:

- **Shortcut Diagnosis:** We provide a diagnostic view of *object-driven shortcuts* in ZS-CAR. We empirically identify strong co-occurrence priors and asymmetric learning difficulty as the root causes of the shortcuts. Furthermore, we expose the limitations of baselines by quantifying how unseen inputs collapse into seen-especially frequent-training pairs, using FSP/FCP.
- **Diagnosis-Driven Framework:** We propose `RCORE` comprising CPR to suppress co-occurrence priors by expanding supervision over originally absent compositions and TORC to promote temporally grounded verb learning.
- **Effectiveness & Robustness:** We demonstrate that `RCORE` is effective, achieving superior performance on unseen compositions over baselines on Sth-com and EK100-com datasets, with multiple VLM backbones. Moreover, through extensive quantitative analyses, we prove that our approach mitigates co-occurrence biases and learns temporally grounded verb features.

2 Related Work

Compositional Action Recognition (CAR). CAR aims to recognize actions by factoring them into verbs and objects. A recurring challenge is learning verb representations that are robust to static cues (*e.g.*, objects/scenes) and capture temporal dynamics [5, 23, 27, 38]. Prior work addresses this by modeling object-level interactions [27, 38] or by training verb representations to be less sensitive to object appearance [5]. Related directions include *open-vocabulary* CAR, which expands the action label space beyond a fixed taxonomy (often via text-driven

semantics) [5,24]. In this work, we study Zero-Shot Compositional Action Recognition (ZS-CAR) under a fixed verb/object vocabulary, where verb and object label sets are shared across splits but some verb–object *pairs are held out* for testing [15,16,19,43]. In ZS-CAR, conditional learning [15,19] and disentanglement-oriented designs [16] improve joint modeling, yet our study highlights that these models can still exhibit *object-driven shortcut* behavior under sparse and skewed compositional supervision (see Section 3, Section 5.2).

Compositional Zero-Shot Learning (CZSL) for image understanding. In the image domain, CZSL focuses on recognizing novel compositions of seen primitives (*e.g.*, attribute–object) under sparse compositional supervision [17,26, 28,29,31,41]. Representative approaches include feature disentanglement [34,45], modular factorization [32], and CLIP-based prompt learning [21,31]. Compared to image CZSL, ZS-CAR additionally requires temporal grounding of verbs, which changes both the failure modes and the regularization opportunities. In contrast, we reveal how sparse compositional supervision and the higher learning difficulty of verbs jointly amplify shortcut reliance in videos, and propose training-time regularizers tailored to ZS-CAR.

Shortcut learning. Models can learn *shortcuts* by exploiting spurious correlations that are easier to fit than the intended evidence [11, 14, 30, 35]. Such shortcuts are closely tied to skewed training distributions and co-occurrence statistics, which can induce reliance on dataset priors [1,37]. In compositional settings, prior work reweights rare combinations to mitigate co-occurrence bias [17]. In video understanding, several works report reliance on static cues such as scenes [2,6,9,39] and objects [3,5] instead of motion-sensitive evidence. Building on these observations, we study a ZS-CAR-specific manifestation: *object-driven shortcuts* where object cues dominate verb prediction under sparse verb–object supervision, and we quantify this behavior with dedicated diagnostics.

Vision-Language Model (VLM). Recent VLMs broadly follow two paradigms: encoder-based models (*e.g.*, CLIP [33] and its video extensions [18, 25, 40, 46]) and decoder-based models (*e.g.*, LLaVA [22], Qwen-VL [4]). In ZS-CAR, however, inference and evaluation operate over a *fixed* verb–object label space with *explicit compositional scores* (*i.e.*, logits over $\mathbb{Y}^V \times \mathbb{Y}^O$), which provides a controlled interface for diagnosis and for designing training-time regularizers that are *transferable across backbones*. Such decoder-based VLM interfaces often do not expose, in a standardized and reproducible way, explicit scores over the fixed label space, as they primarily provide language outputs at inference time. Therefore, we focus on encoder-based VLMs in this controlled setting, and show that object-driven shortcuts persist even with a video-pretrained backbone [40].

3 Diagnosis: Why ZS-CAR Models Fail?

3.1 Zero-Shot Compositional Action Recognition

We study Zero-Shot Compositional Action Recognition (ZS-CAR) [19], where the goal is to recognize *unseen verb–object compositions* that are held out from

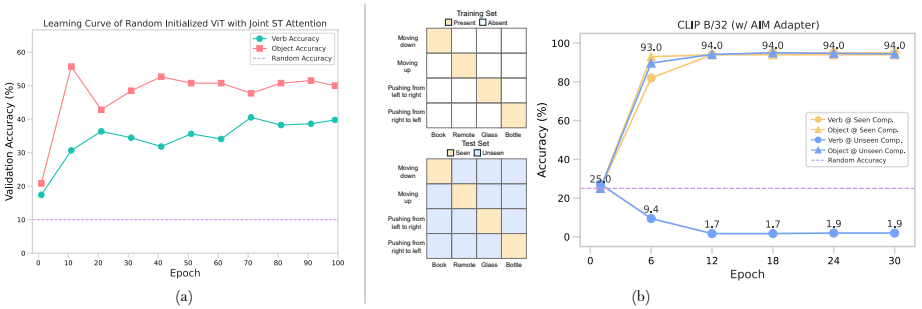


Fig. 2: Controlled experiments demonstrate object-driven shortcut. (a) **Objects are easier than verbs.** A randomly initialized ViT trained on a balanced 10×10 Sth-com subset learns objects substantially faster than verbs. (b) **Co-occurrence bias induces object-driven shortcuts.** On a perfectly biased training split, CLIP (with AIM [42]) achieves high object accuracy, but verb accuracy on bias-conflict unseen compositions drops below chance, indicating shortcut-driven verb failures. Best viewed with zoom and color.

the training set, while the verb and object vocabularies are *fixed* and shared across train/validation/test splits. Thus, unlike open-set or open-vocabulary settings, ZS-CAR does not introduce unseen *verbs* or *objects* at test time—only unseen *compositions* (verb–object pairs). Let $\mathbb{Y}^V$ and $\mathbb{Y}^O$ denote the sets of verb and object labels, and let each composition label be a pair $\mathbf{y}^C = (\mathbf{y}^V, \mathbf{y}^O) \in \mathbb{Y}^V \times \mathbb{Y}^O$. The training set is $\mathbb{D}_{\text{train}} = \{(\mathbf{X}_i, \mathbf{y}_i^C)\}_{i=1}^N$, and its set of observed (seen) compositions is $\mathbb{Y}_{\text{seen}} = \{\mathbf{y}_i^C \mid (\mathbf{X}_i, \mathbf{y}_i^C) \in \mathbb{D}_{\text{train}}\}$. The space of unseen compositions is $\mathbb{Y}_{\text{unseen}} = (\mathbb{Y}^V \times \mathbb{Y}^O) \setminus \mathbb{Y}_{\text{seen}}$. The validation and test sets contain both seen and unseen compositions, enabling evaluation of compositional generalization.

3.2 Diagnostic Metrics

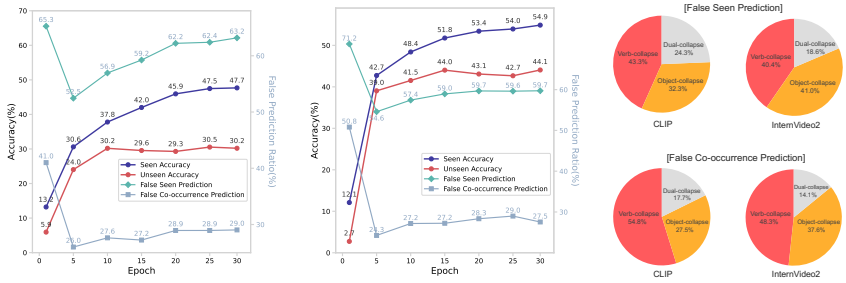
To better understand ZS-CAR model behavior, we introduce two diagnostics: (1) *training-bias* metrics that quantify over-reliance on training co-occurrence statistics, and (2) the *Compositional Gap* that measures the benefit of modeling verb–object dependence beyond independent recognition.

Training-bias metrics. To quantify reliance on co-occurrence priors, we define two misclassification ratios on the *unseen* subset of an evaluation split $\mathbb{Y}_{\text{unseen}}$. *False Seen Prediction (FSP)* is the fraction of misclassified unseen composition samples whose predictions $\hat{\mathbf{y}}^C$ fall into the seen composition categories $\mathbb{Y}_{\text{seen}}$. *False Co-occurrence Prediction (FCP)* is the fraction of misclassified unseen composition samples whose prediction $\hat{\mathbf{y}}^C$ falls into the frequent compositions $\mathbb{Y}_{\text{freq}}$. Let $f(\mathbf{y}^C)$ denote the training frequency of a seen composition $\mathbf{y}^C \in \mathbb{Y}_{\text{seen}}$. We define the set of frequent seen compositions $\mathbb{Y}_{\text{freq}}$ as

$$\mathbb{Y}_{\text{freq}} = \{\mathbf{y}^C \in \mathbb{Y}_{\text{seen}} \mid f(\mathbf{y}^C) > \mu_f + \sigma_f\}, \tag{1}$$

where μ_f and σ_f are the mean and standard deviation of $\{f(\mathbf{y}^C)\}_{\mathbf{y}^C \in \mathbb{Y}_{\text{seen}}}$.

To further analyze shortcut patterns, we decompose FSP/FCP into three component-wise cases based on prediction errors: (i) Verb-collapse, where the



(a) CLIP (b) InternVideo2 (c) FSP/FCP breakdown

Fig. 3: Shortcut reliance correlates with poor unseen generalization. We show learning curves of C2C [19] on Sth-com with CLIP [33] and InternVideo2 [40]. (a) With CLIP, the seen–unseen accuracy gap grows together with FSP/FCP, indicating strong overfitting to seen compositions. (b) Even with the video-pretrained InternVideo2 backbone, the model still shows a notable seen–unseen gap and high FSP. (c) FSP/FCP breakdown at the final epoch: CLIP is dominated by verb-collapse due to object-driven shortcuts, while InternVideo2 exhibits milder but persistent shortcut behavior. Best viewed with zoom and color.

predicted object is correct but the verb is incorrect; (ii) Object-collapse, where the predicted verb is correct but the object is incorrect; and (iii) Dual-collapse, where both verb and object are incorrect. Using the metrics, we can identify whether co-occurrence bias primarily harms verb or object learning.

Compositional Gap. We introduce *Compositional Gap* (Δ_{CG}) to examine whether a model benefits from predicting the joint verb–object composition beyond getting the two components correct independently. Given joint logits over $\mathbb{Y}^V \times \mathbb{Y}^O$, we obtain the predicted composition $\hat{\mathbf{y}}^C = (\hat{\mathbf{y}}^V, \hat{\mathbf{y}}^O)$ by selecting the top-1 composition label. We then compute top-1 accuracies Acc^V, Acc^O, and Acc^C respectively. Then we define Δ_{CG} as

$$\Delta_{\text{CG}} = \text{Acc}^C - (\text{Acc}^V \times \text{Acc}^O). \tag{2}$$

We report Δ_{CG} alongside FSP/FCP and unseen composition accuracy as a *diagnostic* metric, rather than interpreting it as a standalone evaluation criterion.

3.3 Empirical Diagnosis of Object-driven Shortcuts in ZS-CAR

In this section, we empirically diagnose object-driven shortcut behavior in ZS-CAR and analyze how it correlates with weak verb learning and poor generalization to unseen compositions. We show that this behavior persists even with a modern video-pretrained VLM backbone [40], and provide additional evidence in Section 5 and Section S4 (supplementary material).

Objects are easier to learn than verbs. To compare the learning difficulty of verbs and objects, we train a randomly initialized ViT [10] on a controlled subset of Sth-com [19]. Specifically, we construct a balanced 10×10 subset containing all 100 verb–object pairs to isolate learning dynamics from compositional sparsity. Based on prior observations that models tend to fit easier cues earlier

under limited supervision [30, 35], we compare the learning curves of verb *vs.* object prediction. In Figure 2 (a), the model learns objects faster and reaches higher accuracy than verbs under this controlled setting, suggesting that object prediction is easier than verb prediction in our ZS-CAR setup.

Object-driven shortcuts indeed exist in ZS-CAR. Under skewed co-occurrence statistics, models may rely on easier cues as shortcuts instead of learning the intended evidence [3, 6, 11, 20, 30, 35]. We test whether CLIP exhibits such object-driven shortcuts by constructing a *bias-controlled* training split where each verb is strongly correlated with a particular object (see construction details in the supplementary material), and evaluating on two splits: (i) *bias-aligned* seen compositions and (ii) *bias-conflict* unseen compositions. As shown in Figure 2 (b), CLIP achieves high object accuracy *on unseen compositions* while verb accuracy *on unseen compositions* drops below the uniform-chance level ($1/|\mathbb{Y}^V|$), indicating that object cues dominate verb prediction under this biased supervision. In Section S3.1 (supplementary material), we observe this trend even with random initialization, implying shortcuts stem from the training distribution and objective, not just pretraining.

Shortcut reliance correlates with poor unseen generalization. Next, we test whether shortcut reliance correlates with weak unseen generalization in a standard SOTA pipeline. Figure 3 plots validation accuracies together with FSP/FCP for C2C [19] on Sth-com [19]. As shown in Figure 3 (a), with a CLIP backbone [33], increasing FSP/FCP accompanies a widening seen–unseen gap, consistent with growing reliance on co-occurrence statistics. Our error breakdown in Figure 3 (c) further shows that, among unseen samples that collapse to seen pairs, a large fraction corresponds to *verb-collapse*, supporting the presence of object-driven shortcut behavior. As shown in Figure 3 (b) and (c), with an InternVideo2 [40] backbone, the trends are more stable; however, FSP remains high, and the FCP breakdown still exhibits substantial verb-collapse, suggesting that stronger video pre-training alone is insufficient to eliminate the tendency to overfit to seen pairs via co-occurrence-driven shortcuts.

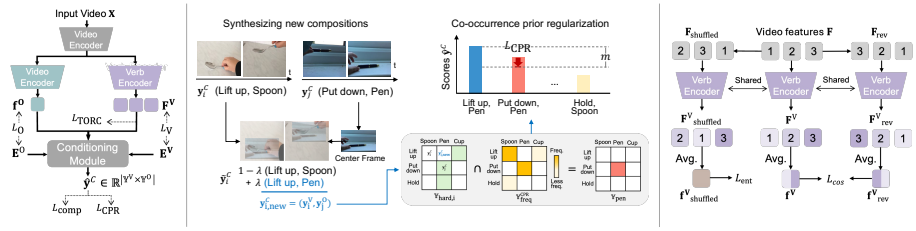
Δ_{CG} indicates degraded compositional behavior on unseen compositions. We report Δ_{CG} of C2C [19] with CLIP [33] and InternVideo2 [40] backbones. As shown in Table 1, the model exhibits negative Δ_{CG} on unseen compositions even with the video-pretrained backbone, indicating that Acc^C falls below the independent reference $\text{Acc}^V \times \text{Acc}^O$. Together with the training-bias metrics (FSP/FCP), these results support that shortcut reliance is closely associated with weak unseen compositional generalization in current ZS-CAR pipelines.

Table 1: Δ_{CG} on Sth-com.

Backbone	Split	Δ_{CG}
CLIP	Seen	+3.24
	Unseen	−0.42
InternVideo2	Seen	+2.24
	Unseen	−0.63

4 RCORE

We introduce RCORE, a diagnosis-driven learning framework that mitigates *object-driven shortcuts* in ZS-CAR. RCORE improves compositional generalization by strengthening verb–object representations under sparse supervision through two



(a) Overview (b) Co-occurrence Prior Regularization (c) Temporal Order Regularization

Fig. 4: Overview of RCORE. (a) Overview of our proposed RCORE framework. (b) CPR synthesizes plausible yet unseen verb–object compositions and penalizes frequently seen hard negatives by a margin-based regularizer. (c) TORC penalizes alignment between original and temporally perturbed feature vectors, enforcing explicit temporal order modeling and reducing object-driven shortcuts.

complementary components: (i) Co-occurrence Prior Regularization (CPR) expands supervision over novel compositions and leverages frequently seen pairs as hard negatives to suppress co-occurrence priors, and (ii) Temporal Order Regularization for Composition (TORC) loss promotes temporally grounded verb learning over static object cues. Following prior work [16, 19], we employ an adapter-based tuning, AIM [42] for CLIP [33] and LoRA [13] for InternVideo2 [40], as our backbone. Figure 4 (a) gives an overview of RCORE.

Feature extraction. Given a video $\mathbf{X}$ with T frames, the backbone encoder outputs frame-level features $\mathbf{F} \in \mathbb{R}^{T\times D}$. We transform them into verb features $\mathbf{F}^V \in \mathbb{R}^{T\times D}$ and object features $\mathbf{f}^O \in \mathbb{R}^D$ using dedicated encoders. We obtain text embeddings for verbs and objects, $\mathbf{E}^V \in \mathbb{R}^{|\mathbb{Y}^V|\times D}$ and $\mathbf{E}^O \in \mathbb{R}^{|\mathbb{Y}^O|\times D}$, by feeding class-specific prompts into the text encoder.

4.1 CPR: Co-occurrence Prior Regularization

To mitigate object-driven shortcuts under sparse and skewed compositional supervision, we propose CPR, a training strategy that injects supervision for *synthesized* verb–object pairs while preserving the stability of closed-world optimization. CPR combines (i) synthesized composition supervision, (ii) a margin-based regularizer that down-weights *frequently* seen hard negatives, and (iii) a batch-adaptive expansion of the composition label space to make these new pairs trainable without optimizing over the full $\mathbb{Y}^V \times \mathbb{Y}^O$ space.

Synthesized composition supervision. To provide explicit supervision for compositions absent from the training set, we first synthesize videos that represent new *composition* labels. Given a training sample $\mathbf{X}_i$ with composition label $\mathbf{y}_i^C = (\mathbf{y}_i^V, \mathbf{y}_i^O)$, we construct a new video $\tilde{\mathbf{X}}_i$ by injecting the static object cue from another video $\mathbf{X}_j$ (with $\mathbf{y}_j^O \neq \mathbf{y}_i^O$) into the high-motion regions of $\mathbf{X}_i$, as shown in Figure 4 (b). Specifically, for each frame $k \in \{1, \ldots, T\}$ we obtain

$$\tilde{\mathbf{X}}_i^k = \left(\mathbf{1} - \lambda\, \mathbf{M}_i^k\right) \odot \mathbf{X}_i^k \;+\; \left(\lambda\, \mathbf{M}_i^k\right) \odot \mathbf{X}_j^{\lfloor T/2 \rfloor}, \tag{3}$$

where $\lambda \in [0, 1]$ controls the injection strength and $\mathbf{M}_i^{(k)} \in \{0, 1\}^{H\times W}$ is the high-motion region mask for frame k, extracted by a learning-free estimator [9].

The operator $\odot$ denotes element-wise multiplication, with $\mathbf{M}_i^k$ broadcast across channels. Accordingly, we use soft labels for the synthesized video as follows:

$$\tilde{\mathbf{y}}_i^O = (1-\lambda)\,\mathbf{y}_i^O + \lambda\,\mathbf{y}_j^O, \tag{4}$$

$$\tilde{\mathbf{y}}_i^C = (1-\lambda)\,\phi(\mathbf{y}_i^V, \mathbf{y}_i^O) + \lambda\,\phi(\mathbf{y}_i^V, \mathbf{y}_j^O), \tag{5}$$

where $\phi : \mathbb{Y}^V \times \mathbb{Y}^O \to \{0,1\}^{|\mathbb{Y}^V \times \mathbb{Y}^O|}$ maps a composition $(\mathbf{y}^V, \mathbf{y}^O)$ to its one-hot vector in the flattened joint label space.

Co-occurrence prior regularization loss. Even after adding supervision for synthesized compositions, skewed co-occurrence statistics can still make the model over-score a small set of *frequently* seen pairs. To counter this score collapse, we enforce a margin constraint that makes the *target synthesized* composition $\mathbf{y}_{i,\text{new}}^C = (\mathbf{y}_i^V, \mathbf{y}_j^O)$ outrank *frequent seen hard negatives* by at least m. We define the hard-negative label set for $\mathbf{y}_{i,\text{new}}^C$ as follows:

$$\mathbb{Y}_{\text{hard},i} = \left\{ (\mathbf{y}^V, \mathbf{y}^O) \;\middle|\; \left[(\mathbf{y}^V = \mathbf{y}_i^V) \oplus (\mathbf{y}^O = \mathbf{y}_j^O)\right] \wedge (\mathbf{y}^V, \mathbf{y}^O) \neq (\mathbf{y}_i^V, \mathbf{y}_i^O) \right\}, \tag{6}$$

where $\oplus$ denotes the exclusive or operator. Let $\mathbb{Y}_{\text{freq}}^{\text{CPR}} \subset \mathbb{Y}_{\text{seen}}$ be the set of frequently seen compositions (see Section S3.2 of the supplementary material for the exact criterion). Then we define the CPR loss with margin m over the penalty set $\mathbb{Y}_{\text{pen}} = \mathbb{Y}_{\text{hard},i} \cap \mathbb{Y}_{\text{freq}}^{\text{CPR}}$ as:

$$L_{\text{CPR}} = \sum_{\mathbf{y}^C \in \mathbb{Y}_{\text{pen}}} \max\Big(0,\; s(\mathbf{y}^C) - s(\mathbf{y}_{i,\text{new}}^C) + m\Big), \tag{7}$$

where $s(\mathbf{y}^C)$ denotes the logit assigned to the composition label $\mathbf{y}^C$.

Batch-adaptive label-space expansion. Conventional ZS-CAR pipelines optimize classification only over *seen* compositions $\mathbb{Y}_{\text{seen}}$ (closed-world over compositions), which makes it non-trivial to directly supervise *novel* verb–object pairs during training. On the other hand, naively applying cross-entropy over the full $\mathbb{Y}^V \times \mathbb{Y}^O$ space can harm unseen accuracy by treating most unseen compositions as negatives throughout training (see Section 5.4). To provide an effective supervision for novel compositions, we introduce a *batch-adaptive label space expansion* strategy. For each mini-batch, we construct an expanded label set as $\mathbb{Y}_{\text{exp}} = \mathbb{Y}_{\text{seen}} \cup \mathbb{Y}_{\text{new}}$, where $\mathbb{Y}_{\text{new}}$ denotes the label set for the newly synthesized videos. We compute cross-entropy only over $\mathbb{Y}_{\text{exp}}$. This preserves the optimization stability of closed-world training while progressively injecting supervision for new compositions.

4.2 TORC: Temporal Order Regularization for Composition

To counter object-driven shortcuts in verb learning, we introduce TORC, a temporal *order-sensitivity* regularizer. While CPR expands supervision over compositions, verb recognition can still collapse to static object cues due to the asymmetric learning difficulty between verbs and objects. As shown in Figure 4 (c),

TORC directly regularizes the verb representation to depend on temporal structure: it (i) discourages invariance to temporal reversal by separating forward and reversed features, and (ii) penalizes over-confident verb predictions when temporal order is disrupted by encouraging high-entropy outputs. Together, these constraints reduce reliance on static cues and promote temporally grounded verb representations.

Temporal perturbation for regularization. Given frame-level features $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)$, we form two regularization views: (i) a reversed sequence $\mathbf{F}_{\text{rev}} = (\mathbf{f}_T, \ldots, \mathbf{f}_1)$, and (ii) a temporally shuffled sequence $\mathbf{F}_{\text{shuffled}} = \pi(\mathbf{F})$, where π is a random permutation sampled from $\mathcal{P}$. We then feed $\mathbf{F}_{\text{rev}}$ and $\mathbf{F}_{\text{shuffled}}$ into the verb encoder to obtain perturbed verb features $\mathbf{f}^V_{\text{rev}} \in \mathbb{R}^D$ and $\mathbf{f}^V_{\text{shuffled}} \in \mathbb{R}^D$.

TORC loss. We first push the model to distinguish forward and reversed temporal semantics. We minimize the cosine similarity between the original verb $\mathbf{f}^V$ and the reversed feature vectors $\mathbf{f}^V_{\text{rev}}$: $L_{\cos} = \frac{\mathbf{f^V}^\top \mathbf{f}^V_{\text{rev}}}{\|\mathbf{f}^V\| \|\mathbf{f}^V_{\text{rev}}\|}$. Temporal reversal often flips action meaning (*e.g.*, opening *vs.* closing), yet the state-of-the-art model [19] yields a high cosine similarity (0.92; see Figure S7 of the supplementary material), indicating weak temporal discrimination.

Next, we regularize the model to avoid confident verb predictions when temporal structure is disrupted. Given the verb text embeddings $\mathbf{E}^V = \{\mathbf{e}^V_m\}_{m=1}^{|\mathbb{Y}^V|}$, we define the negative entropy loss $L_{\text{ent}} = \sum_{m=1}^{|\mathbb{Y}^V|} p_m \log p_m$ with the predicted probability p_m of the m-th verb class based on the cosine similarity between visual and textual features as:

$$p_m = \frac{\exp(\cos(\mathbf{f}^V_{\text{shuffled}}, \mathbf{e}^V_m)/\tau)}{\sum_{n=1}^{|\mathbb{Y}^V|} \exp(\cos(\mathbf{f}^V_{\text{shuffled}}, \mathbf{e}^V_n)/\tau)}, \tag{8}$$

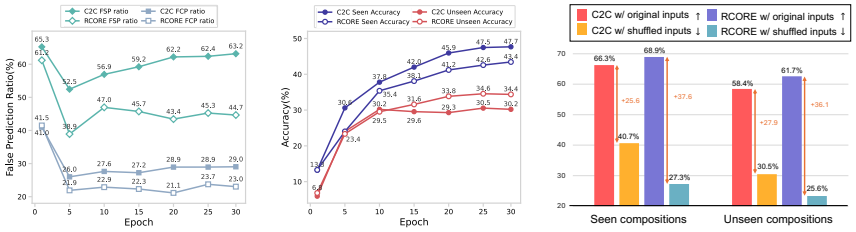
where τ is a temperature parameter. Maximizing entropy prevents the model from inferring verbs from static spatial cues alone, enforcing reliance on temporal evidence. We define the TORC loss as: $L_{\text{TORC}} = L_{\cos} + L_{\text{ent}}$.

4.3 Training

Component loss. We compute verb logits $\mathbf{S}_V \in \mathbb{R}^{|\mathbb{Y}^V|}$ and object logits $\mathbf{S}_O \in \mathbb{R}^{|\mathbb{Y}^O|}$ based on the cosine similarity scores. We apply standard cross-entropy losses for verb and object prediction: $L_{\text{com}} = L_V + L_O$.

Composition loss. Following prior work [19], we obtain the composition logits $\mathbf{S}_C \in \mathbb{R}^{|\mathbb{Y}^V \times \mathbb{Y}^O|}$ by aggregating factorized logits: $\mathbf{S}_C = \mathbf{S}_V \boxplus \mathbf{S}_{O|V} + \mathbf{S}_O \boxplus \mathbf{S}_{V|O}$, where $\mathbf{S}_{O|V}$ and $\mathbf{S}_{V|O}$ represent the conditional logits and $\boxplus$ denotes broadcasted addition into the joint verb–object space. Then we employ the cross-entropy loss integrated with our batch-adaptive label space expansion (Section 4.1). Given $\mathbf{s}_c \in \mathbf{S}_C$, the final composition loss utilizing the dynamically constructed denominator $\mathbb{Y}_{\text{exp}}$ is formulated as: $L_{\text{comp}} = -\log(\exp(\hat{s}_c)/\sum_{c=1}^{|\mathbb{Y}_{\text{exp}}|} \exp(s_c))$.

Total loss. We train RCORE with the combined objective L_{total} defined as $L_{\text{total}} = \alpha L_{\text{com}} + \beta L_{\text{comp}} + \gamma L_{\text{TORC}} + \delta L_{\text{CPR}}$, where α, β, γ and δ are hyperparameters.



(a) (b) (c)

Fig. 5: Analysis on the effects of RCORE on the Sth-com [19] dataset. (a) RCORE suppresses the growth of FSP/FCP during training, unlike the baseline. (b) This reduces the seen–unseen gap and improves unseen composition accuracy. (c) On the temporal subset, RCORE shows a larger drop under temporal shuffling, indicating stronger temporal dependence over static cues. Best viewed with zoom and color.

5 Experimental Results

Building on our diagnosis, we evaluate whether RCORE reduces shortcut reliance and improves compositional generalization. We first describe the experimental setup (Section 5.1), then present diagnostic analyses (Section 5.2), followed by quantitative comparisons (Section 5.3), and finally ablation studies (Section 5.4).

5.1 Experimental setup

Limitations of conventional evaluation. Existing ZS-CAR works [15,16,19] rely on the *closed-world* protocol, where inference is restricted to compositions that appear in the validation/test split. This protocol hides a model's tendency to over-predict seen compositions and obscures co-occurrence-driven behaviors. More critically, *test-set–tuned* bias calibration selects a bias term using test labels, inflating unseen accuracy and undermining fair comparison.

Our evaluation setup. To expose genuine generalization, we adopt an *open-world, unbiased* protocol by default: (i) inference spans the full space $\mathbb{Y}^V \times \mathbb{Y}^O$, and (ii) no test labels are used to tune biases. We also report closed-world H.M. and AUC for comparison with existing methods. We show open-world biased results using a validation set-tuned bias in Table S8 (supplementary material).

Datasets. We evaluate on two ZS-CAR benchmarks: Sth-com [19] and our curated EK100-com. Sth-com is derived from Something-Something V2 [12] and Something-Else [27], containing 79K videos with 161 verbs and 248 objects. Following prior work [36, 44], we also evaluate on the *Temporal* subset consisting of verbs requiring temporal reasoning. We additionally introduce the **EPIC-KITCHENS-100-composition (EK100-com)** dataset, constructed by repurposing EK100 [7] following the Sth-com protocol. EK100-com includes 71K egocentric videos, 81 verbs, and 216 objects. Due to the long-tailed label distribution of EK100, it exhibits *severe compositional sparsity*. Its label coverage ratio is only 7.5%, which is considerably lower than the 12.8% observed in Sth-com [19]. Full details are in the supplementary material (Section S2).

Evaluation metrics. We report top-1 composition, verb and object accuracies. Verb and object accuracies are conditioned on the *seen/unseen compo-*

sition: *verb@seen-comp* and *object@seen-comp* are computed on samples with $\mathbf{y}^C \in \mathbb{Y}_{\text{seen}}$, while *verb@unseen-comp* and *object@unseen-comp* are computed on samples with $\mathbf{y}^C \in \mathbb{Y}_{\text{unseen}}$. We report the corresponding harmonic mean (H.M.) between seen-comp and unseen-comp accuracies for composition, verb, and object. We also report the compositional gap Δ_{CG} (2). For a detailed illustration of our evaluation setting, please refer to the supplementary material (Section S1).

Baselines. We use two encoder-based VLMs, CLIP-B/16 [33] and InternVideo2-CLIP-B/14 [40], as backbones for ZS-CAR methods. We also report the InternVideo2-1B [40] backbone results in the supplementary material. We primarily compare RCORE against the current SOTA, C2C [19] and Jung et al. [16]; since the code for the latter's CLIP variant is not publicly available, we report our re-implementation that reproduces its closed-world performance. We also compare against LogicCAR [43], a recent ZS-CAR method, in the closed-world setting only, as its code is not publicly available. Since it builds on C2C (enhance) [19], a variant of C2C [19] with additional training constraints, we apply RCORE to the same enhanced backbone (denoted as RCORE (enhance)) and report its results for a fair comparison. To isolate the effect of conditional learning, we further evaluate independent modeling baselines, 'AIM' [42] (for CLIP) and 'LoRA' [13] (for InternVideo2), named after their respective adapter tuning methods.

5.2 Analysis

We empirically show that RCORE reduces reliance on co-occurrence priors and mitigates object-driven shortcuts in verb learning, yielding more robust compositional behavior than the baseline [19].

RCORE less relies on co-occurrence prior. In Figure 5 (a), we present the False Seen Prediction (FSP) and False Co-occurrence Prediction (FCP) curves of RCORE and the baseline [19] on Sth-com. While the baseline's FSP and FCP increase substantially during training (53% → 63% and 26% → 29%), those of RCORE remain consistently lower and even decrease (47% → 44% and 23% → 23%), indicating markedly less dependence on co-occurrence statistics.

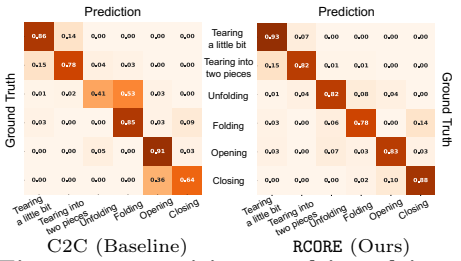


Fig. 6: RCORE mitigates object-driven shortcuts in verb learning. We visualize confusion matrices for six representative verbs on *unseen* compositions of the Sth-com [19] test set. All values are normalized frequencies across the six verb classes.

Improved unseen generalization. In Figure 5 (b), we present the learning curve of our RCORE and the baseline [19] on Sth-com. RCORE significantly improves unseen composition accuracy upon the baseline (34% vs. 30%), and reduces the seen–unseen accuracy gap (9 points vs. 17 points).

The learned verb representations are indeed effective. Inspired by prior work [44], we evaluate models on the *Temporal* subset of Sth-com [19]. Specifically, to measure a model's dependence on static cues, we compare its performance when using temporally shuffled verb features $\mathbf{f}^V_{\text{shuffled}}$ versus the original

Table 2: Sth-com [19] results. We show the top-1 verb, object, and composition classification accuracy (%). We report performance on both seen and unseen compositions, along with their harmonic mean (H.M.).

Backbone	Method	Verb			Object			Composition		
		@Seen Comp	@Unseen Comp	H.M.	@Seen Comp	@Unseen Comp	H.M.	Seen ($\Delta_{CG}\uparrow$)	Unseen ($\Delta_{CG}\uparrow$)	H.M.
CLIP	AIM [42]	54.99	39.19	45.76	67.19	53.43	59.53	40.24 (+3.29)	18.50 (−2.44)	25.35
	C2C [19]	63.60	54.36	58.62	67.72	56.10	**61.36**	46.31 (+3.24)	30.08 (−0.42)	36.47
	Jung et al. [16]	66.57	53.55	59.35	66.56	50.90	57.69	**48.41** (+4.10)	25.25 (−2.01)	33.19
	RCORE	65.73	59.00	**62.18**	64.79	56.34	60.27	44.99 (+2.40)	**33.90** (+0.66)	**38.67**
InternVideo2	LoRA [13]	48.83	36.50	41.78	64.79	58.61	61.55	32.83 (+1.19)	19.71 (−1.68)	24.63
	C2C [19]	70.35	63.29	66.63	70.03	63.44	**66.58**	**51.50** (+2.24)	39.53 (−0.63)	44.73
	RCORE	71.65	66.65	**69.06**	67.96	64.56	66.22	50.20 (+1.51)	**43.98** (+0.95)	**46.88**

features $\mathbf{f}^V$. As shown in Figure 5 (c), across both seen and unseen compositions, RCORE exhibits a substantially larger performance gap between original and shuffled features. This demonstrates that its verb representations are effectively grounded in temporal dynamics. In Figure 6, we visualize verb confusion matrices on *unseen* compositions of the Sth-com test set. Unlike C2C [19], which frequently confuses opposite verbs (*e.g.*, 'unfolding' vs. 'folding'), RCORE discriminates them well, demonstrating that it effectively mitigates object-driven shortcuts in verb learning.

5.3 Comparisons with State-of-the-Art

Sth-com. As shown in Table 2, RCORE achieves clear gains on Sth-com [19] under a rigorous open-world setting. Compared to the C2C [19] baseline, RCORE improves the verb@unseen-comp by 4.6 points and the unseen composition accuracy by 3.8 points with the CLIP [33] backbone, and by 3.4 points and 4.5 points with InternVideo2 (IV2) [40] backbone, respectively. These gains are consistent with stronger verb recognition observed in our analysis (Section 5.2). Notably, RCORE yields a positive compositional gap (Δ_{CG}) on unseen compositions, whereas all baselines remain negative, indicating improved compositional behavior on unseen pairs. Jung et al. [16] yields the lowest unseen composition accuracy among the CLIP baselines while showing the highest seen composition accuracy. Compared to C2C [19], it improves seen composition accuracy by 2.1 points but degrades unseen composition accuracy by 4.8 points, reflecting a strong bias toward seen compositions and limited open-world generalization. In Table 3, RCORE achieves the highest H.M. and AUC within each backbone group under the closed-world, test-set-tuned protocol used in prior work [16, 19, 43].

Table 3: Closed-world results on Sth-com [19].

Method	Closed-world	
	H.M.	AUC
C2C [19]	42.8	23.8
Jung et al. [16]	42.6	23.5
RCORE	**43.2**	**25.3**
C2C (enhance) [19]	44.5	26.0
LogicCAR [43]	45.2	27.0
RCORE (enhance)	**46.1**	**27.5**

EK100-com. Despite the strong co-occurrence prior in the EK100-com training compositions, RCORE achieves higher unseen composition accuracy and the best H.M. As shown in Table 4, compared to C2C [19], RCORE substantially improves unseen composition accuracy—by 6.9 points with the CLIP [33] backbone and

Table 4: EK100-com results. We show the top-1 verb, object, and composition classification accuracy (%). We report performance on both seen and unseen compositions, along with their harmonic mean (H.M.).

Backbone	Method	Verb			Object			Composition		
		@Seen Comp	@Unseen Comp	H.M.	@Seen Comp	@Unseen Comp	H.M.	Seen ($\Delta_{CG}\uparrow$)	Unseen ($\Delta_{CG}\uparrow$)	H.M.
CLIP	AIM [42]	59.97	39.99	47.98	56.23	46.76	51.06	38.58 (+4.86)	14.98 (−3.71)	21.58
	C2C [19]	65.52	49.57	56.44	56.54	46.99	51.32	**42.70** (+5.65)	21.56 (−1.73)	28.65
	Jung et al. [16]	63.38	47.63	54.39	55.85	42.12	48.02	39.50 (+4.10)	16.96 (−3.10)	23.73
	RCORE	66.19	54.31	**59.66**	54.35	49.88	**52.02**	39.70 (+3.73)	**28.41** (+1.32)	**33.12**
InternVideo2	LoRA [13]	59.73	32.99	42.51	51.08	39.34	44.45	36.43 (+5.92)	5.63 (−7.35)	9.75
	C2C [19]	68.95	57.01	62.41	60.54	52.31	56.12	**47.30** (+5.55)	30.33 (+0.51)	36.96
	RCORE	67.76	59.78	**63.52**	56.28	56.00	**56.14**	41.03 (+2.89)	**37.31** (+3.83)	**39.08**

7.0 points with the IV2 [40] backbone—despite a decrease in seen composition accuracy. This pattern reflects reduced reliance on co-occurrence statistics and improved recognition of unseen compositions. Moreover, RCORE yields a positive Δ_{CG} on unseen compositions (+1.32 with CLIP and +3.83 with IV2), which is consistent with improved compositional behavior under highly sparse supervision. In the case of Jung et al. [16], the model yields lower composition accuracy than C2C [19] on both seen (−3.2 points) and unseen (−4.6 points) compositions with the CLIP [33] backbone, suggesting that its benefit may be limited under the highly sparse compositional supervision of EK100-com.

5.4 Ablation Studies

Effects of RCORE components. In Table 5 (a), using TORC alone yields the largest gain in verb generalization, improving *verb@unseen-comp* by 3.5 points, and gain in unseen composition accuracy by 3.8 points, compared to the baseline without CPR and TORC (C2C [19]). Using CPR alone primarily improves unseen composition accuracy by 3.1 points, while reducing seen composition accuracy 4.2 points, indicating a seen–unseen trade-off. Combining CPR and TORC achieves the best overall results, improving unseen composition accuracy by +5.0 points and composition H.M. by +2.7 points, suggesting complementary benefits of mitigating co-occurrence bias and enforcing temporal sensitivity.

Effects of TORC components. Table 5 (b) evaluates the roles of L_{cos}, which enforces discrimination of opposite temporal semantics, and L_{ent}, which prevents reliance on static cues. Each term individually improves performance, while using both terms yields the largest gains in the verb@unseen-comp (+3.5 points) and unseen composition accuracy (+3.8 points).

Effects of label space in CPR. In Table 5 (c), we validate the batch-adaptive label space expansion in CPR. Compared to closed-world training over $\mathbb{Y}_{seen}$, our batch-adaptive label space improves unseen composition accuracy (+2.3 points) and H.M. by injecting supervision for synthesized compositions while avoiding unstable optimization. In contrast, naively optimizing cross-entropy over the full $\mathbb{Y}^V \times \mathbb{Y}^O$ space severely hurts unseen composition accuracy by treating most unseen compositions as negatives throughout training.

Table 5: Ablation study on the Sth-com [19] validation set. We use C2C [19] with the CLIP [33] backbone as our baseline, using the unbiased open-world setting. We report performance on both seen and unseen compositions, along with their harmonic mean (H.M.). V@S and V@U denotes verb@seen-comp and verb@unseen-comp, while O@S and O@U denotes object@seen-comp and object@unseen-comp accuracies.

(a) Effects of RCORE components.

CPR	TORC	Verb		Object		Composition		
		V@S	V@U	O@S	O@U	Seen	Unseen	H.M.
		65.26	54.02	67.21	56.81	47.31	30.14	36.83
✓		63.95	54.05	63.97	59.52	43.09	33.22	37.51
	✓	64.93	57.49	66.20	57.23	46.28	33.92	39.15
✓	✓	67.06	58.13	63.67	58.66	45.00	35.16	**39.48**

(b) Effects of TORC components.

L_{cos}	L_{ent}	Verb		Object		Composition (OW)		
		V@S	V@U	O@S	O@U	Seen	Unseen	H.M.
		65.26	54.02	67.21	56.81	47.31	30.14	36.83
✓		64.15	55.96	67.71	55.97	47.13	31.11	37.48
	✓	64.84	54.82	67.57	56.23	47.46	30.84	37.39
✓	✓	64.93	57.49	66.20	57.23	46.28	33.92	**39.15**

(c) Effects of label space in CPR.

Label Space	Verb		Object		Composition		
	V@S	V@U	O@S	O@U	Seen	Unseen	H.M.
Closed-world ($\mathbb{Y}_{\text{seen}}$)	64.61	54.26	67.21	57.46	46.78	30.80	37.14
Full Open-world ($\mathbb{Y}^V \times \mathbb{Y}^O$)	67.36	45.01	68.49	52.29	50.72	17.34	25.85
Batch-adaptive ($\mathbb{Y}_{\text{exp}}$)	64.17	54.55	63.60	58.70	42.74	33.05	**37.28**

(d) Effects of penalty set in CPR.

Penalty Set $\mathbb{Y}_{\text{pen}}$ Choice	Verb		Object		Composition		
	V@S	V@U	O@S	O@U	Seen	Unseen	H.M.
None (w/o L_{CPR})	64.17	54.55	63.60	58.70	42.74	33.05	37.28
All Hard Negatives ($\mathbb{Y}_{\text{hard}}$)	61.95	57.22	47.99	49.05	31.85	30.82	31.33
Frequent Hard Negatives ($\mathbb{Y}_{\text{hard}} \cap \mathbb{Y}^{\text{CPR}}_{\text{freq}}$)	65.36	55.86	63.20	57.62	43.60	33.16	**37.67**

Effects of penalty set in CPR. In Table 5 (d), we study the effect of the penalty set $\mathbb{Y}_{\text{pen}}$. Compared to disabling L_{CPR}, penalizing *frequent* hard negatives ($\mathbb{Y}_{\text{hard}} \cap \mathbb{Y}^{\text{CPR}}_{\text{freq}}$) yields the best trade-off, improving unseen composition accuracy and achieving the highest H.M. of 37.7%. In contrast, penalizing all hard negatives degrades object prediction and hurts overall composition performance, suggesting that restricting penalties to frequent confounders is critical.

6 Conclusions

In this work, we revisit zero-shot compositional action recognition and identify *object-driven shortcuts* as a key failure mode. We attribute this behavior to compositional sparsity and asymmetric verb–object learning difficulty. Using our diagnostic metrics, we show that existing models overfit to co-occurrence priors, which degrades verb learning and thereby harms generalization to unseen compositions. We propose RCORE comprising two components: CPR, which expands supervision to synthesized compositions, and TORC, which enforces temporal-order sensitivity. Across two benchmarks under an open-world evaluation protocol without test-tuned bias calibration, RCORE consistently improves performance on unseen compositions, moving ZS-CAR toward more reliable verb–object compositional reasoning.

Acknowledgements

This work was supported by the NAVER Cloud Corporation and partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea Government(MSIT) under grant RS-2024-00353131 (25%) and Information & Communications Technology Planning & Evaluation(IITP)-ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT) (IITP-2025-RS-2023-00259004, 25%). Additionally, it was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (RS-2025-02216217, 50%).

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: CVPR (2018) 4
2. Bae, K., Ahn, G., Kim, Y., Choi, J.: DEVIAS: Learning disentangled video representations of action and scene for holistic video understanding. In: ECCV (2024) 1, 4
3. Bahng, H., Chun, S., Yun, S., Choo, J., Oh, S.J.: Learning de-biased representations with biased representations. In: ICML (2020) 4, 7
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025) 4
5. Chatterjee, D., Sener, F., Ma, S., Yao, A.: Opening the vocabulary of egocentric actions. In: Adv. Neural Inform. Process. Syst. (2023) 3, 4
6. Choi, J., Gao, C., Messou, J.C., Huang, J.B.: Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In: Adv. Neural Inform. Process. Syst. (2019) 4, 7
7. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. IJCV **130**, 33–55 (2022) 1, 2, 3, 11
8. Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Van Gool, L.: Large scale holistic video understanding. In: ECCV (2020) 1
9. Ding, S., Li, M., Yang, T., Qian, R., Xu, H., Chen, Q., Wang, J., Xiong, H.: Motion-aware contrastive video representation learning via foreground-background merging. In: CVPR (2022) 4, 8
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 6
11. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: ICLR (2018) 2, 4, 7
12. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The "something something" video database for learning and evaluating visual common sense. In: ICCV (2017) 1, 11
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022) 8, 12, 13, 14
14. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: NeurIPS (2019) 4
15. Jiang, D., Jing, H., Ma, Y., Zheng, N.: Beyond image classification: A video benchmark and dual-branch hybrid discrimination framework for compositional zero-shot learning. In: CVPR (2025) 1, 4, 11
16. Jung, H., Bak, J.H., Jeong, Y., Lee, G., Ahn, J., Kim, E.S.: Zero-shot compositional video learning with coding rate reduction. In: ICCV (2025) 1, 4, 8, 11, 12, 13, 14
17. Kim, H., Lee, J., Park, S., Sohn, K.: Hierarchical visual primitive experts for compositional zero-shot learning. In: ICCV (2023) 4

18. Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., Qiao, Y.: Unmasked teacher: Towards training-efficient video foundation models. In: ICCV (2023) 4
19. Li, R., Feng, Z., Xu, T., Li, L., Wu, X.J., Awais, M., Atito, S., Kittler, J.: C2c: Component-to-composition learning for zero-shot compositional action recognition. In: ECCV (2024) 1, 2, 3, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15
20. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: ECCV (2018) 2, 7
21. Li, Y., Liu, Z., Chen, H., Yao, L.: Context-based and diversity-driven specificity in compositional zero-shot learning. In: CVPR (2024) 4
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Adv. Neural Inform. Process. Syst. (2023) 4
23. Liu, Y., Liu, F., Jiao, L., Bao, Q., Li, S., Li, L., Liu, X.: Knowledge-driven compositional action recognition. Pattern Recognition **163**, 111452 (2025) 3
24. Luo, Z., Ghosh, S., Guillory, D., Kato, K., Darrell, T., Xu, H.: Disentangled action recognition with knowledge bases. In: North American Chapter of the Association for Computational Linguistics (2022) 4
25. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval. In: ACM MM (2022) 4
26. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Open world compositional zero-shot learning. In: CVPR (2021) 4
27. Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., Darrell, T.: Something-Else: Compositional action recognition with spatial-temporal interaction networks. In: CVPR (2020) 1, 2, 3, 11
28. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: CVPR (2017) 4
29. Naeem, M.F., Xian, Y., Tombari, F., Akata, Z.: Learning graph embeddings for compositional zero-shot learning. In: CVPR (2021) 4
30. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: De-biasing classifier from biased classifier. In: NeurIPS (2020) 2, 4, 7
31. Nayak, N.V., Yu, P., Bach, S.: Learning to compose soft prompts for compositional zero-shot learning. In: ICLR (2023) 4
32. Purushwalkam, S., Nickel, M., Gupta, A., Ranzato, M.: Task-driven modular networks for zero-shot compositional learning. In: ICCV (2019) 4
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 4, 6, 7, 8, 12, 13, 14, 15
34. Saini, N., Pham, K., Shrivastava, A.: Disentangling visual embeddings for attributes and objects. In: CVPR (2022) 4
35. Scimeca, L., Oh, S.J., Chun, S., Poli, M., Yun, S.: Which shortcut cues will DNNs choose? a study from the parameter-space perspective. In: ICLR (2022) 2, 4, 7
36. Sevilla-Lara, L., Zha, S., Yan, Z., Goswami, V., Feiszli, M., Torresani, L.: Only time can tell: Discovering temporal data for temporal modeling. In: WACV (2021) 11
37. Singh, K.K., Mahajan, D., Grauman, K., Lee, Y.J., Feiszli, M., Ghadiyaram, D.: Don't judge an object by its context: learning to overcome contextual bias. In: CVPR (2020) 2, 4
38. Sun, P., Wu, B., Li, X., Li, W., Duan, L., Gan, C.: Counterfactual debiasing inference for compositional action recognition. In: ACM MM (2021) 3
39. Wang, J., Gao, Y., Li, K., Lin, Y., Ma, A.J., Cheng, H., Peng, P., Huang, F., Ji, R., Sun, X.: Removing the background by adding the background: Towards

background robust self-supervised video representation learning. In: CVPR (2021) 4
40. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Wang, C., Chen, G., Pei, B., Zheng, R., Xu, J., Wang, Z., et al.: Internvideo2: Scaling video foundation models for multimodal video understanding. ECCV (2024) 2, 4, 6, 7, 8, 12, 13, 14
41. Wu, P., Lai, Q., Fang, H., Xie, G.S., Yin, Y., Lu, X., Wang, W.: A conditional probability framework for compositional zero-shot learning. In: ICCV (2025) 4
42. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: AIM: Adapting image models for efficient video understanding. In: ICLR (2023) 5, 8, 12, 13, 14
43. Ye, G., Li, L., Li, K., Xiao, J., Chen, L.: Zero-shot compositional action recognition with neural logic constraints. In: ACM MM (2025) 1, 4, 12, 13
44. Yun, S., Kim, J., Han, D., Song, H., Ha, J.W., Shin, J.: Time is matter: Temporal self-supervision for video transformers. In: ICML (2022) 11, 12
45. Zhang, T., Liang, K., Du, R., Chen, W., Ma, Z.: Disentangling before composing: Learning invariant disentangled features for compositional zero-shot learning. TPAMI **47**(2), 1132–1147 (2024) 4
46. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., Hornung, R., Schroff, F., Yang, M.H., Ross, D.A., Wang, H., Adam, H., Sirotenko, M., Liu, T., Gong, B.: VideoPrism: A commercial-grade video foundation model. In: ICML (2024) 4