Implicit Affordance Acquisition via Causal Action–Effect Modeling in the Video Domain

Hsiu-Yu Yang
University of Stuttgart
hsiu-yu.yang@ims.uni-stuttgart.de

Carina Silberer
University of Stuttgart
carina.silberer@ims.uni-stuttgart.de

Abstract

Affordance knowledge is a fundamental aspect of commonsense knowledge. Recent findings indicate that world knowledge emerges through large-scale self-supervised pretraining, motivating our exploration of acquiring affordance knowledge from the visual domain. To this end, we augment an existing instructional video resource to create the new Causal Action–Effect (CAE) dataset[1] and design two novel pretraining tasks—Masked Action Modeling (MAM) and Masked Effect Modeling (MEM)—promoting the acquisition of two affordance properties in models: behavior and entity equivalence, respectively. We empirically demonstrate the effectiveness of our proposed methods in learning affordance properties. Furthermore, we show that a model pretrained on both tasks outperforms a strong image-based visual–linguistic foundation model (FLAVA) as well as pure linguistic models on a zero-shot physical reasoning probing task.

1 Introduction

Affordances refer to the potential actions and interactions that objects offer/are available to intelligent agents (Gibson, 1977). For example, a chair affords sitting and a cup affords drinking. In the context of artificial intelligence, affordance understanding involves training an agent to recognize and interpret affordances in the real world, allowing for seamless action anticipation and planning (Ardón et al., 2021; Nagarajan and Grauman, 2020). Recent advancements in large language models (LLMs) have enabled researchers to use their extensive everyday knowledge to decompose high-level natural language instructions into low-level actions for embodied agents (Ichter et al., 2022; Huang et al., 2022). The lack of physical grounding, however, limits these models to understand affordances (Bisk et al., 2020a). The main challenge in learning affordance is multiplicity. Put simply, different objects of distinct semantic classes potentially facilitate the same action, and a single object can offer multiple feasible actions (Lu et al., 2022). As illustrated in Figure 1(b) & (c), both the “ball” and “pizza cutter” can afford the action “roll” while garlic supports actions like “slide” and “cut”. More precisely, two principles serve as the foundation for realizing multiplicity (Şahin et al., 2007). The first is *behavior equivalence*, which states that different actions can produce the same effect on a given object, e.g., “slice”, “cut” and “chop” can all make “garlic” into smaller pieces (Fig. 1(b)). The second, *entity equivalence*, suggests that executing the same action on different objects can lead to identical outcomes, e.g., rolling an “apple” producing a similar motion change to rolling the “ball” (Fig. 1(c)). Learning such associations poses a significant challenge for generalizing knowledge to novel objects and unfamiliar scenarios (Lu et al., 2022).

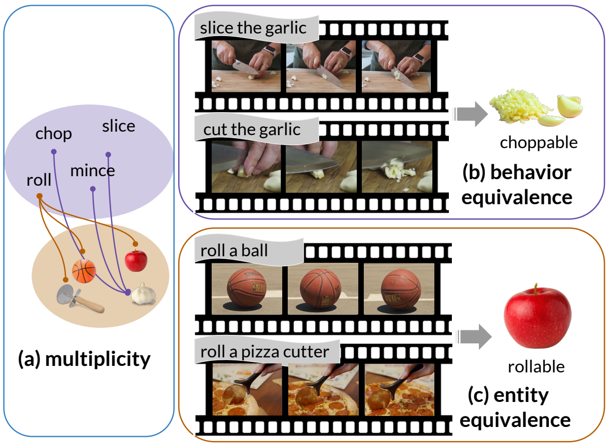


Figure 1: We address the (a) multiplicity issue of affordance learning by modeling causal action–effect during pretraining to implicitly induce two essential properties of affordance: (b) behavior and (c) entity equivalence.

Existing methods address the generalization problem either by identifying shared characteristics among objects of an affordance category, or

[1] The project code can be found at https://github.com/Mallory24/cae_modeling

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 846–871
November 1–4, 2023. ©2023 Association for Computational Linguistics

iteratively reinforcing cross-modal representation consistency (Lu et al., 2022; Luo et al., 2021). However, these approaches are devised for specific affordance tasks in supervised settings, raising questions about their adaptability to other tasks (Bisk et al., 2020b; Aroca-Ouellette et al., 2021). Given recent findings on the emergence of world knowledge from large-scale self-supervised pretraining (Petroni et al., 2019) and the importance of the visual domain for distilling certain knowledge (Shwartz and Choi, 2020; Paik et al., 2021), we study the implicit acquisition of affordance knowledge through visual–linguistic (VL) pretraining. As we seek for a high diversity in action and effect categories, we explore affordance learning in a noisy real-world video domain (Ebert et al., 2023). To this end, we leverage step-by-step instructional videos accompanied by subtitles (Miech et al., 2019), from which we extract affordance-relevant clip–subtitle pairs. The resulting dataset, referred to as CAE(**C**ausal **A**ction–**E**ffect Dataset), contains 4.1M clip–subtitle pairs for modeling diverse actions and their effects. We then introduce two pretraining tasks—MAM and MEM—to induce behavior equivalence and entity equivalence, respectively. To validate our approach, we conduct intrinsic evaluations addressing two research questions: **(RQ1)** can models trained with MAM and MEM adequately learn fundamental principles of affordances? **(RQ2)** what are the benefits of joint task training? We then assess the encoded affordance knowledge in the grounded representations to answer the third question: **(RQ3)** how effective is our causal action–effect pretrained model on solving an affordance probing task?

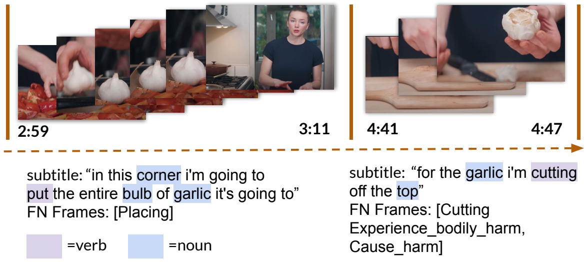
2 Related Work

Affordance Learning. Several works mine affordances from text corpora by identifying semantically plausible verb–object pairs (Loureiro and Jorge, 2018; Persiani and Hellström, 2019; Chao et al., 2015). Closely related are selectional preferences (Pantel et al., 2007; Erk, 2007), i.e., typical arguments (e.g., objects) of a verbal predicate. Another line of research targets *visual affordances*, their detection and categorization in visual input (Hassanin et al., 2022). Affordances underlie the multiplicity property, i.e., multiple objects can be mapped to one affordance category and vice versa, making supervised learning challenging. Few works address this by enhancing the multimodal representation consistency iteratively (Lu et al., 2022) or by finding joint object features of a given affordance class (Luo et al., 2021). Inspired by the robotics domain (Şahin et al., 2007; Dağ et al., 2010; Dehban et al., 2016; Jaramillo-Cabrera et al., 2019), we model the relationships between actions, objects, and the observed effects for encoding affordance knowledge implicitly via a self-supervised setup. In contrast to Merullo et al. (2022) that model object trajectories as effects in a closed simulated environment, we use action–object–effect relations mined from diverse web-crawled instructional videos.

Causal Modeling of Action Verbs. Grounding the meaning of action verbs to the state changes of the manipulated objects is gaining attention in the NLP community (Sampat et al., 2022b). Gao et al. (2016) are the first to explore causal verb modeling in the cooking domain (Regneri et al., 2013) for grounded semantic role labeling. Bosselut et al. (2017) use symbolic action–effect modeling for procedural text understanding. Gao et al. (2018) introduce multimodal action–effects prediction—linking action verbs to their effects in static images. Several works approach action–effect modeling in a simulated environment. Zellers et al. (2021) collect action and object state transitions in AI2-THOR (Kolve et al., 2017), and ground language models to the physical world through symbolic representations (e.g., `isWarm=True`). Hanna et al. (2022) address effect prediction in AI2-THOR in a more challenging setup where the image showing the post-action is to be chosen, while Dagan et al. (2023) predict change labels from the visual input. Our goal, in contrast, is to implicitly augment models with affordance knowledge through causal action–effect modeling. Moreover, we do not bound object states to a fixed set of categorical labels, but exploit the temporal dimension to represent perceptual effect changes.

3 The CAE Dataset

We base our work on the hypothesis that affordance knowledge can emerge from large-scale self-supervised pretraining. We choose to build upon HowTo100M (Miech et al., 2019), the largest and most diverse instructional video dataset available at the time of writing (Tang et al., 2019; Zhukov et al., 2019; Kuehne et al., 2019a). It contains 136M weakly-paired clip–subtitles, varying across domains. As our interest lies in modeling



subtitle: "in this corner i'm going to put the entire bulb of garlic it's going to"
FN Frames: [Placing]

subtitle: "for the garlic i'm cutting off the top"
FN Frames: [Cutting Experience_bodily_harm, Cause_harm]

Figure 2: CAE Clip–subtitle pairs.

causal action–effect, we employ a series of automatic procedures to extract useful video clips from HowTo100M: (1) identify a set of result verbs by leveraging various linguistic resources (2) locate casual action–effect video clips via parsing subtitles to match the result verb set. We call our resulting clip–subtitle pairs the CAE dataset.[1]

3.1 Identify Result Verbs

To model perceptual causal change of actions, we are interested in a specific verb type called result verbs. Result verbs cause state changes on their arguments, including changes in volume, area, gradable scale, and motion (Levin and Malka Rappaport, 2010). One can reuse an existing collection of result verbs from Gao et al. (2018); yet, it has limited coverage (62 verb classes, of which 39 are covered in our list). Furthermore, to facilitate a reproducible and systematic method for identifying result verbs, we define 2 criteria that a potential result verb should meet: (1) **visualness**: it can be visually perceivable (2) **effect-causing**: it can cause physical results on the object it acts upon. We use several semantic resources, including VerbNet (Kipper et al., 2006) for its informative selectional restrictions, and imSitu (Yatskar et al., 2016), to leverage its exhaustive frame–semantic annotations on visual verbs. Moreover, we employ FrameNet (FN; Baker et al., 1998) to extract (undisambiguated) situational frames a verb could evoke and use the associated frame elements to identify potential result verbs.[2] We will further use the extracted frames for generalization analysis on unseen verb classes (detailed in Sec. 5). To automate result verb identification, we define heuristics to verify **visualness** and **effect-causing** properties.

Visualness. We retain potential result verbs from imSitu's visually perceivable verbs, whose semantic role in the second position is valid, e.g. `Item`.[3] From VerbNet, we extract visual verbs based on the selectional restrictions on the thematic role of its verb class. Precisely, if a verb class, e.g., *spank-18.3*, specifies the `Patient` role to be concrete or solid, we consider its grouped verb members to possess visualness, e.g., "whisk" and "whip".

	Videos	Clips	Top 5 Result Verbs
food	167k	2.15M	make, put, cook, mix, cut
hobbies	60k	0.8M	make, put, cut, pull, fold
cars	12k	0.1M	make, put, pull, turn, push
pets	8k	90k	make, put, cut, set, build
sports	3k	39k	make, put, cut, pull, build

Table 1: Number of unique videos, video clips, and the top 5 result verbs across 5 selected video domains.

Effect-causing. To automatically determine the effect-causing characteristic of a verb, we check if the thematic role combination of its verb class on VerbNet follows (`Agent`, `Patient`, `Result`), e.g., the verb "split" in the verb class *break-45.1* is confirmed. We then cross-check with FN on the potential result verbs that passed the previous visualness test and the initial effect-causing test to ensure this property. Specifically, we check if a candidate result verb can evoke any frame that contains either the `Result` or the `Effect` frame element, e.g., the verb "simmer" evokes the APPLY_HEAT frame, and has the `Result` element.

To consolidate the **visualness** and the **effect-causing** information obtained from the various lexical resources, we merge the verbs on their lemma, e.g.,"grill" is judged as a result verb because its verb sense *grill-45.3* captures visualness and effect-causing properties, even though the other, *grill-26.3-2*, lacks such information on VerbNet. We derive in total **236** sure cases with both **visualness** and **effect-causing** characteristics (an overview of sure and unsure cases can be found in Tab. 7 of App. A.1), which we use to locate causal action–effect video clips described in the following.

3.2 Extract CAE Video Clips

Given the large size of the HowTo100M dataset, we down-sample the video pool with several heuristics and focus on 13 video domains with a high density of unique result verbs (see App. A.2 for details). To extract video clips from this pool relevant for causal action–effect modeling, we rely on

[2] See App. A.1 for details on the versions we use.

[3] A list of invalid roles are detailed in App. A.1

the paired subtitles and their linguistic information such as PoS tags and dependency labels. Particularly, we only keep clips with a single occurrence of one of our results verbs. To mitigate information leakage due to overlapping video frames, we enforce a minimum difference of 5 seconds between adjacent clips at the per-video level.[4] In addition, to annotate proxy objects (i.e., nouns), we find the set of objects (within a subtitle) by considering the intersection of nouns labeled with `dobj` or `pobj` relations and with high concreteness ratings (i.e., > 4; Brysbaert et al., 2014). Through this process, we derive in total **4.1**M video clips with **235** unique result verb types.[5] Figure 2 gives examples (the comparison to the underlying HowTo100M clips can be found in App. A.2, Fig. 5).[6] As Table 1 shows, most of the verbs are prevalent across domains, e.g. "make". Since our goal is to maintain a naturalistic dataset, we keep such common verbs (see App. A.2 for details.). Recall that we built upon noisy weakly-paired clip–text data, resulting in many items without visual occurrences of both our target actions and objects (70% of 288 CAE samples analyzed by a postgraduate student). We keep these "background examples" as a preliminary study (Kuehne et al., 2019b).

4 Causal Action–Effect Modeling

Our goal is to induce the basic principles of affordance knowledge, i.e., behavior equivalence and entity equivalence (Fig. 1 of Sec. 1). We design two pretraining tasks: MAM and MEM, and further explore the benefit of joint task training by alternating task-specific samples to optimize the model, referred to as Multi-task on Causal Action–Effect Modeling (MULTI-CAE). In order to gauge the understanding of affordance properties, the corresponding intrinsic tasks called Masked Action Prediction (MAP) and Masked Effect Prediction (MEP) are introduced. We build upon the existing video–language hierarchical framework HERO (Li et al., 2020), consisting of a **Cross-Modal Transformer** to learn contextualized embeddings between a subtitle and a video clip locally and a **Temporal Transformer** to learn video embeddings on the global context (i.e, the whole video clip). Figure 3 illustrates the overall architecture and the pretraining tasks (for model details see App. A.4).

4.1 Masked Action Prediction (MAP)

Task Definition. Our underlying assumption is that a video clip of the CAE dataset visually depicts the change of the pre-condition (`[BEF]`), the action process (`[ACT]`), and the post-condition (`[AFT]`) in sequential order throughout an action execution (Sampat et al., 2022b). Given a CAE clip–subtitle pair, where the action $a \in A$ (a result verb) is masked. The task is to predict verb a.

Masked Action Modeling (MAM). We address MAP on the local context, i.e., the contextualized multimodal embeddings computed by the Cross-Modal Transformer (Fig. 3 (a) MAM). To prepare the multimodal inputs, we extract a series of video frames $V = \{\mathbf{v}_j\}_{j=1}^{|V|}$ from the video clip every 2 seconds, and tokenize the corresponding subtitle into a sequence of tokens $S = \{\mathbf{s}_i\}_{i=1}^{|S|}$ (refer to App. A.4 for details). We replace the (masked) target verb with the special `[MASK]` token.

Training. During pretraining, we seek to encourage the model to learn holistic semantics. We thus mask the verb a and each word in S with a chance of 15%.[7] We empirically verify the benefit of this masking strategy during preliminary experiments (see App. A.5 for details). The objective is to reconstruct $\mathbf{s}_{a*}$ (the action verb and some random words) based on the observation of unmasked tokens $S_{\backslash \mathbf{s}_{a*}}$ and the video clip V such that the learnable parameters θ are updated with the loss function:

$$\mathcal{L}_{MAM}(\theta) = -\log P_\theta(\mathbf{s}_{a*}|S_{\backslash \mathbf{s}_{a*}} \| V)$$

Inference. During inference, only the target verb a is masked in the input. We feed the preprocessed input to HERO, and consider the token with the highest logits in the MAM head's output as the predicted token $\hat{a}$.

4.2 Masked Effect Prediction (MEP)

Task Definition. Recall from MAP (Sec. 4.1), the effect of an action on the object is assumed to be visually perceivable in a CAE clip–subtitle pair. The goal of MEP is to predict the masked `[AFT]` (post-condition) subclips in a discriminative setup:

[4] We discard consecutive clips as they usually capture redundant information, i.e., the same result verb.

[5] Only the phrasal verb *warm_up* is not considered.

[6] To align with our visual feature extraction procedure, the shown timestamps are extended 3 seconds before the original timestamp (see App. A.4). Note the released CAE dataset contains the original timestamp.

[7] Following BERT, 80% and 15% of the target verbs are replaced with `[MASK]` and a random token, respectively, 5% of them are unchanged.

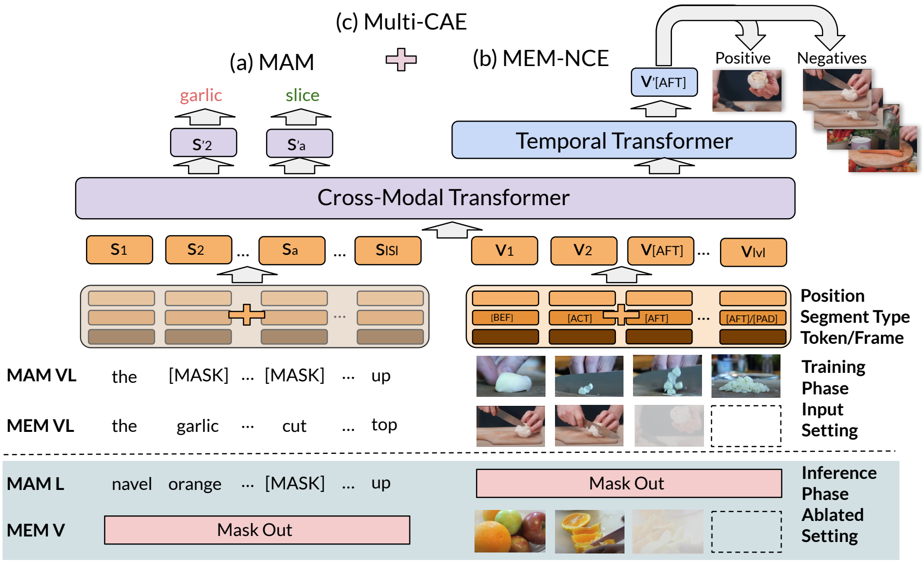


Figure 3: Our proposed pretraining tasks on the hierarchical video–language architecture HERO (Li et al., 2020): (a) MAM for behavior equivalence on the output of Cross-Modal Transformer (Sec. 4.1) (b) MEM for entity equivalence on the output of Temporal Transformer (Sec. 4.2). (c) MULTI-CAE for joint-task training. During inference (blue-gray background), we also ablate one modality via attention masks for a systematic intrinsic evaluation.

for each `[AFT]` subclip (frame), the correct frame from candidate frames with the same **video id** as the given video clip is to be selected.[8]

Masked Effect Modeling (MEM). Unlike MAP, the task is addressed globally with the temporal contextualized embeddings computed by the Temporal Transformer (Fig.3 (c) MEM). Given a preprocessed pair of video frames $V = \{\mathbf{v}_j\}_{j=1}^{|V|}$ and subtitle tokens $S = \{\mathbf{s}_i\}_{i=1}^{|S|}$, we arbitrarily divide the video clip into three parts in a near-equal manner: `[BEF]` (pre-condition), `[ACT]` (action), `[AFT]` (post-condition)[9] and mask the video frames $\mathbf{v}_{\texttt{[AFT]}}$ corresponding to the `[AFT]` post-condition subclips with zero-vectors.

Training. During pretraining, the objective is to reconstruct $\mathbf{v}_{\texttt{[AFT]}}$ given the observation of unmasked video frames $V_{\setminus \mathbf{v}_{\texttt{[AFT]}}}$ and subtitle S:

$$\mathcal{L}_{MEM}(\theta) = -\log P_\theta(\mathbf{v}_{\texttt{[AFT]}} | V_{\setminus \mathbf{v}_{\texttt{[AFT]}}} \| S)$$

Concretely, following Li et al. (2020), we optimize the model with a contrastive loss by employing the softmax version of the NCE loss function (Józefowicz et al., 2016), thus,

$$P_\theta(\mathbf{v}_{\texttt{[AFT]}} | V_{\setminus \mathbf{v}_{\texttt{[AFT]}}} \| S) \approx \frac{\exp\left(\mathbf{v}'_{\texttt{[AFT]}} \cdot \mathbf{v}_{\texttt{[AFT]}}/\tau\right)}{\sum_{i=0}^{C} \exp\left(\mathbf{v}'_{\texttt{[AFT]}} \cdot c_i/\tau\right)}$$

where τ is a temperature to control the strength of the penalties on the negative samples, and where the global contextualized embeddings $\mathbf{v}'_{\texttt{[AFT]}}$, computed by the Temporal Transformer, are taken to compute the similarity to each video frame c_i in C, a candidate set comprising the correct $\mathbf{v}_{\texttt{[AFT]}}$ and negative video frames sampled from the same **video id** as the input video clip V.[10]

Inference. Identical to training, we mask post-condition video frames $\mathbf{v}_{\texttt{[AFT]}}$ for reconstruction. By using the trained MEM head to compute the dot product between $\mathbf{v}'_{\texttt{[AFT]}}$ and each of the candidate frames from the same video id, the highest score obtained is considered as the model's prediction.

5 Experiments

We first evaluate our proposed methodology for multimodal affordance learning intrinsically on MAP and MEP, and then on the task-specific probing benchmark Physical Reasoning about Objects through Space and Time (PROST).

[8] The rest the of intra-video clip `[AFT]` frames are excluded from the candidate set (see App. A.6 for details).

[9] We leave automatic action localization (Zhang et al., 2022) to determine effect frames more reliably, to future work.

[10] In a preliminary study on various negative sampling strategies, we found this strategy most beneficial for teaching fine-grained effect changes (see App. A.6).

	video clips	% zero-shot videos	actions	objects
train	2,818,433	–	–	–
val	646,423	1.07	45.45	0.90
test	644,986	1.10	45.47	0.86
all	4,109,842	–	–	–

Table 2: Number of video clips and % of zero-shot samples w.r.t. the train set in terms of unique video ids, action (verb) classes, object (noun) types.

5.1 Intrinsic Evaluation Setup

The intrinsic evaluation (see Sec. 4.1 & 4.2 for the task descriptions) addresses **(RQ1)** & **(RQ2)** (see Sec.1). Precisely, we assess these two affordance principles by **(1)** quantifying their generalization ability in a zero-shot framework, i.e., evaluate how well a model performs on unseen action (verb) and object (noun) classes; **(2)** examining the generalization ability from the perspective of event (Baker et al., 1998) and lexico-taxonomic (Miller, 1994) semantics. We thus implement an algorithm controlling the splitting of the CAE dataset.[11]

Dataset Split. To examine the action generalization ability on the level of event semantics in a **generalized zero-shot** setting (Xian et al., 2017), we assign verbs either to seen or unseen classes within a FrameNet frame and remove the restriction that all the verb classes of the test set should be unseen. In total, there are **143** seen and **92** unseen verb classes, resp. As shown in Table 2, the train/dev/test split ratio follows 70%/15%/15% and nearly 50% of instances in the dev/test set contain unseen actions (verbs) (see App. A.3 for relevant frequency statistics for verbs and verb–noun combinations in train/test). Note that the test set is not manually annotated by humans and serves solely as a reference set (Kuehne et al., 2019a).

Model Ablations. To ensure a faithful assessment of the generalization ability, we diverge from Li et al. (2020) and use randomized weights rather than RoBERTa-B's pretrained weights to initialize HERO's Cross-Modal Transformer.[12] The models pretrained with MAM, MEM, and MULTI-CAE tasks are referred to as **MAM-VL**$_{Rnd}$, **MEM-VL**$_{Rnd}$, and **MULTI-CAE-VL**$_{Rnd}$, respectively. Moreover, to see if the modalities provide complementary information,[13] we train model variants with one ablated modality for comparison: **MAM-L**$_{Rnd}$ and **MEM-V**. We do this through zero-masking on the inter-attention mask. In other words, **MAM-L**$_{Rnd}$ and **MEM-V** are tasked to reconstruct the masked textual/visual target tokens under their own respective unimodal input (see Appendix A.7 Ablated Models). During inference, we perform a sanity check, where we ablate one modality to see whether the linguistic/visual context is sufficient to solve the task as illustrated in Figure 3. See Appendix A.7 for the pretraining details and the hyperparameters we adopted.

Metrics. For MAP, we report the macro-average accuracy on seen/unseen verb classes and their harmonic mean (Xian et al., 2017), along with the micro-average accuracy. For the MEP task, we report the micro-average accuracy, which measures the correctness of predicting all masked [AFT] frames per instance.

5.2 Probing Task: PROST

To address **(RQ3)** (Sec. 1), we assess the encoded affordance knowledge in our causal action–effect pretrained models by performing a zero-shot evaluation on the PROST task (Aroca-Ouellette et al., 2021). It intends to probe models on physical commonsense knowledge in a textual cloze-style format, including 6 affordance concepts: **stackable, rollable, graspable, breakable, slidable** and **bounceable** for 38 object classes in total. Each of them comprises two templates: (1) the **original template** asks the model to choose the object among 4 objects that affords a given action, while (2) its **inverse** asks for the object that cannot afford the action; e.g., *eggs* is the non-stackable object in [*eggs, books, blocks, boxes*]. The distribution of **correct answer** positions is uniform across the test instances. Therefore, a model that is found robust exhibits similar effectiveness on the original and its inverse template, as well as on all answer positions.

Models. We compare pure language models (LM) against variants of MAM-L, MAM-VL, and MULTI-CAE-VL, whose textual encoders are initialized with the pretrained RoBERTa-B model (Hugging Face, Wolf et al., 2020) before CAE pretraining. Additionally, to test our hypothesis that the video domain better captures action–

[11] The algorithm is reproducible, allowing researchers to customize their own splits based on their requirements.

[12] According to Li et al. (2020), RoBERTa's pretrained weights (12 layers) are partially taken to initialize the Cross-Modal Transformer (6 layers) of HERO.

[13] The action is verbally mentioned (see Sec. 3.2.)

Test Set Input	Model	Seen	Unseen	HM	Micro	FN frame	Co-Hypo
Multimodal	Majority baseline	0.7	1.1	2.0	14.6	-	-
	Multi-CAE-VL$_{Rnd}$	23.6	13.5	17.2	30.9	16.6	36.1
	MAM-VL$_{Rnd}$	**26.8**	**14.9**	**19.1**	**31.7**	**17.5**	34.5
Language	RoBERTa-B †	7.6	5.2	6.2	13.3	3.7	8.8
	Multi-CAE-VL$_{Rnd}$	15.1	10.2	12.2	19.6	11.7	**39.0**
	MAM-VL$_{Rnd}$	16.5	10.6	12.9	19.7	12.6	36.6
	MAM-L$_{Rnd}$	18.4	14.4	16.2	23.3	10.2	37.3

Table 3: **Left**: Results (%) for MAP on (1) Multimodal (video + subtitles) and (2) Language (subtitles) test input for Seen/Unseen verbs. HM = Harmonic Mean, Micro = micro-average accuracy. RoBERTa-B † is a linguistic baseline. **Right**: Proportion (in %) of false predictions on the set of unseen verb classes, which generalizes in terms of event and lexico-taxonomic semantics, measured through a shared FrameNet (FN) frame and direct Co-Hyponymy, respectively. Best results in **boldface**, best result in each input mode underlined.

effects due to the temporal dimension, we compare against the image-based visual–linguistic foundation model FLAVA (Singh et al., 2022), known for its strong unimodal and multimodal representations. Note that before undergoing multimodal pretraining, the textual encoder of FLAVA is initialized with unimodal pretrained weights,[14] making its linguistic understanding comparable to pretrained text-based models. The models thus all have some prior linguistic knowledge, allowing us to focus on the effect of CAE pretraining. Regarding the LMs, we report results for RoBERTa-B, and the average (coined AvgLM) of the best results that Aroca-Ouellette et al. (2021) report for each of their LMs (GPT, GPT2, BERT, RoBERTa, ALBERT V2).[15]

For inference, all models are fed *textual input* only, and we follow Aroca-Ouellette et al.'s (2021) procedure to obtain the probabilities of each candidate, based on the logits of pretrained MLM (RoBERTa-B & FLAVA) and MAM heads (ours).

6 Results

6.1 Intrinsic Tasks: MAP

The accuracy of all models on predicting the correct verb is overall lower on the **language-input** test set compared against the **multimodal** set, where they are fed both, subtitles and video clips (Block 2 vs. 1, Tab. 3, left). The best model, MAM-VL$_{Rnd}$, drops by -10.3pp accuracy on seen and -4.3pp on unseen verb classes, suggesting that the task and test set is not trivially solvable without visuals. It also indicates that the VL model has successfully learned to ground verbs by leveraging the visual effect change (cf. Sampat et al., 2022a). Indeed, when omitting the `[AFT]` frames (the visual effect) during inference, the accuracy of MAM-VL$_{Rnd}$ drops on Seen and Unseen (-4.5pp and -2.3pp, resp., no table shown). But the benefit of the visual modality for the model seems still limited regarding its generalization ability—the difference on Unseen between MAM-VL$_{Rnd}$ and its ablated variant MAM-L$_{Rnd}$ (pretrained on subtitles), is minimal ($+.5$pp for VL); Multi-CAE-VL$_{Rnd}$ even underperforms MAM-L$_{Rnd}$ on Unseen irrespective of the test input.

Analysis: Generalization Ability. We analyzed the effect of our pretraining tasks on the generalization ability of verbs on the level of situations/events, i.e., semantic frames, and less specific verb senses: We measured the number of cases, in which the reference verb $\mathbf{v}$ and the predicted verb $\tilde{\mathbf{v}}$, $\mathbf{v} \neq \tilde{\mathbf{v}}$, share the same FrameNet (FN) frame or the same direct WordNet (WN) hypernym (co-hyponymy).[16] Table 3 (right) shows that MAM-VL$_{Rnd}$ and Multi-CAE-VL$_{Rnd}$ have the highest proportion of cases for which they did not predict $\mathbf{v}$, but a shared FN frame or a co-hyponym. Moreover, MAM-L$_{Rnd}$ falls short against both VL models across the board on FN frames, in contrast to our findings above on MAP on verb *types*. E.g., MAM-VL$_{Rnd}$ predicted "fry" instead of reference "roast", but both evoke the Apply_heat frame and are co-hyponyms of "cook", while MAM-L$_{Rnd}$ predicted "put" (Placing; "move"; see also Fig. 12 of App. A.5). Noteworthy is also that MAM-VL$_{Rnd}$ is best on frame-level generalization *with* visual input during inference, while Multi-CAE-VL$_{Rnd}$ is the best model on the co-hyponymy relation *without* visual input. The patterns indicate that, first,

[14] The textual encoder is pretrained using the MLM objective on CCNews and BookCorpus.

[15] Following Aroca-Ouellette et al. (2021), we exclude the not directly comparable models T5 (it does not cover **slide**), and UnifiedQA (fine-tuned on task-specific text data).

[16] We used the NLTK toolkit (Bird and Loper, 2004).

Test Set Input	Model	Accuracy
Multimodal	Random baseline	0.27
	MULTI-CAE-VL$_{Rnd}$	59.2
	MEM-VL$_{Rnd}$	**59.9**
Video	MULTI-CAE-VL$_{Rnd}$	58.6
	MEM-VL$_{Rnd}$	57.2
	MEM-V	**59.7**

Table 4: Results (%) for MEP on (1) Multimodal (video clips + subtitles) and (2) Video (video clips) test input.

the visual modality is beneficial for frame semantics, and second, **visual–linguistic (VL) action learning fosters event knowledge, while VL action–effect learning fosters lexico-taxonomic knowledge.**

Qualitative Analysis. Through introspection of individual verb classes, we found the **visually perceivable effect to be beneficial for the majority** of our examined classes. For example, the accuracy of MAM-VL$_{Rnd}$ on Seen "whip" and "bake" drops significantly (−24pp and −20.7pp, resp.), when it is tested with language input only; examples for Unseen are "crumple", and "wring" (−9.1pp, −6.3pp, resp.). However, **for several classes the linguistic context alone gives a strong cue**. "Manicure", e.g., often co-occurs with the word "nail", and "sharpen" with "knife" (we refer to App. A.5, Fig. 10 for an example). For an example of extrapolations of visual effect similarities for capturing behavior equivalence, see Appendix A.5, Fig. 11.

6.2 Intrinsic Tasks: MEP

As shown in Table 4, the accuracy of the multimodal and vision-only model, MEM-VL$_{Rnd}$ and MEM-V, is comparable on visual effect inference on their respective input modes (59.9% and 59.7%). This indicates that linguistic knowledge priors do not contribute much to visual action–effect comprehension. Comparing these models with a variant whose language encoder is initialized with RoBERTa-B before multimodal pretraining supports this—it yields a neglectable difference in accuracy (60% and 59.6% on VL and V test input, resp., no table shown). On the other hand, on video-only inference, MULTI-CAE-VL is slightly more effective than MEM-VL$_{Rnd}$ (+1.4pp). Thus, a linguistic encoder that is trained jointly on both, visual effect and *linguistic* action prediction (MULTI-CAE-VL) may benefit *video-only* effect inference. We refer to Figures 14 and 15 for examples of successful entity equivalence reasoning, and failures due to ambiguous reference.

6.3 Probing Task: PROST

Pretraining models on both, *visual* action *and* effect prediction in the video domain, is beneficial for learning affordance knowledge. As Table 5 (left) shows, MULTI-CAE-VL, which is additionally trained to observe the action process *and* its perceptual effect on objects obtains the best Macro Average (32%), with a difference to AvgLM of up to +19.1pp (**slide**). Notably, it achieves better results on average compared to FLAVA (+16.2pp). This indicates that modeling visual action–effects in the temporal dimension plays a crucial role in effectively encoding latent affordance concepts. Note that **slide** is the only affordance that is contained in our CAE training data, yet, even on **slide**, MAM-VL yields a higher accuracy (+6pp, Tab. 5) than its variant trained only on textual training items (subtitles), MAM-L.

Beneficial is moreover pretraining towards visual effect prediction (i.e., MEM), as the comparison of MULTI-CAE-VL against MAM-VL shows (+5.5pp on average). The former significantly outperforms all other models also on two affordances that are not covered by CAE's train split (**stack** (34%) and **bounce** (38%)). That is, MULTI-CAE-VL learns to extrapolate to unseen actions.
Our models are at least comparable to the pure LMs on all individual affordances except for **break** (RoBERTa-B, 31%). When compared to FLAVA, our models perform slightly worse in the case of **grasp**, though the performance difference is not significant. Introspection showed the performance discrepancy between the affordances is not due to seen vs. unseen, *objects*, in fact, all are seen (cf. Tab. 15, App. A.8). However, we found that the errors made by our model in **break** and **grasp** are related to the frequency of objects in the CAE training set. In other words, our models tend to pick the most common seen object among the 4 choices as its prediction, e.g., wrongly selecting *sugar* as the graspable object.

Model Robustness. We test the models for robustness to the order of answer choices and to template inversion (difference in accuracy on affordance/non-affordance), allowing deeper insights into the models' proper affordance knowledge and language understanding ability, respec-

Model	Stack	Roll	Grasp	Break	Slide	Bounce	Macro	1	2	3	4
MAM-L	30.0	20.0	24.0	27.0	24.0	23.0	24.7	22.5	22.5	22.3	22.3
MAM-VL	22.0	25.0	33.0	26.0	30.0	23.0	26.5	25.0	25.0	25.2	25.1
Multi-CAE-VL	**34.0***	24.0	30.0	26.0	**40.0***	**38.0***	**32.0***	27.4	27.4	27.6	27.7
FLAVA	19.0	26.9	**34.5**	25.5	23.8	24.0	25.5	25.8	21.5	21.0	21.7
RoBERTa-B	27.4	24.9	22.8	**31.0***	26.6	25.5	26.4	45.8	19.4	15.1	15.0
AvgLM †	29.1	**29.0**	28.4	30.7	20.9	20.4	26.3	31.4	25.2	8.5	38.6

Table 5: PROST: **Left**: Results (accuracy %) across six affordances. **Right**: Position accuracy across the correct answer's position. The more balanced, the more robust. †Results taken from PROST (Aroca-Ouellette et al., 2021). Statistically significant differences ($p < 0.05$)[17] are indicated with *.

Model	Stack ↓	Roll ↓	Grasp ↓	Break ↓	Slide ↓	Bounce ↓	Macro Average ↓
MAM-L	60.0	20.0	32.0	14.0	40.0	34.0	33.3
MAM-VL	4.0	30.0	62.0	**12.0**	20.0	6.0	22.3
Multi-CAE-VL	68.0	32.0	56.0	**12.0**	80.0	76.0	54.0
FLAVA	4.5	45.2	68.6	41.9	39.6	31.7	38.6
RoBERTa-B	**0.1**	**14.5**	**5.2**	51.9	17.5	**2.4**	15.3
AvgLM	11.4	14.7	12.7	17.8	**12.8**	9.6	**13.2**

Table 6: PROST: Absolute difference in accuracy between the original template and its inverse across six affordances. The lower the better the model's true question understanding.

tively. Table 5 (right) shows that all our models display the most balanced effectiveness across different answer positions, indicating robustness against syntactic change of the answer position, i.e., the order in which the correct (non-)affordable object is presented in the context. In contrast, text-based models like RoBERTa-B and AvgLM are significantly affected, while FLAVA slightly favors the first answer position.

In terms of robustness to template inversion, as shown in Table 6, Multi-CAE-VL, the overall most effective model in accuracy, is the least balanced, it has on average a 54pp higher accuracy on inverses. Our second overall best model, MAM-VL, is by far more robust (avg. 22.3pp diff.) and is better than FLAVA. This result is somewhat surprising since FLAVA is expected to have a stronger linguistic understanding due to its additional pretraining on textual corpora. **The pure language models turn out to be most robust to inverses** (avg. 13.2pp and 15.3pp for AvgLM and RoBERTa-B, resp.), indicating a higher ability of *true language* understanding.[18] See Appendix A.8, Table 16 for the breakdown of the individual results on original/inversed templates.

7 Discussion

The intrinsic results suggest that both modalities contribute beneficially towards our goal of encoding behavior and entity equivalence. Crucially, our results on the PROST affordance probing task strongly support our joint action–effect modeling in the video domain: Multi-CAE-VL is the most effective model. It is also most robust in terms of language-only (esp. PROST) and visual-only input (MEP) inference modes. We hypothesize that Multi-CAE-VL is more encouraged in taking both modalities into account due to its training objectives on both, vision-conditioned *linguistic* action prediction (MAM) and language-conditioned *visual* effect prediction (MEM).

8 Conclusions

We explore the acquisition of affordance properties–behavior and entity equivalence through large-scale causal action–effect pretraining in the video domain. To this end, we augment an existing instructional video dataset and introduce two pretraining tasks, MAM & MEM. Our empirical results show the successful incorporation of these foundamental properties of affordance. Furthermore, the joint-task-trained model outperforms linguistic counterparts on an affordance probing task. Future work would benefit from a cleaner dataset to explicitly examine the contribution of action/effect modeling, and further exploration of multi-task training.

[18] Recall our CAE pretraining data has noisy subtitles including automatically transcribed speech, which may hinder language understanding.

9 Limitations

In this work, we seek to validate the induction of two affordance properties through large-scale self-supervised visual–linguistic pretraining. It is important to acknowledge that affordance knowledge encompasses a range of concepts (Zhu et al., 2015; Xu et al., 2022), and our assessment is currently limited to a linguistic probing task. Future work is necessary to investigate whether the representation derived from causal action–effect modeling can effectively address other affordance downstream tasks, e.g., purpose-driven affordance understanding (Luo et al., 2021; Zhai et al., 2022). Furthermore, it is worth investigating the potential applications of the encoded latent affordance representations in areas like procedural tasks (Zhou et al., 2023) and robot learning (Bahl et al., 2023). When considering the use of the video domain for large-scale pretraining, one has to remark on certain shortcomings compared to a controllable embodied environment. Firstly, the temporal misalignment of the video clips and subtitles poses a challenge for cross-modal learning (Zhukov et al., 2019; Miech et al., 2020). Moreover, many subtitles consist of incomplete sentences, as most of them are automatically transcribed based on an ongoing speech within fixed time intervals (Kuehne et al., 2019b), restricting the linguistic capabilities of models trained on such data. Secondly, it is computationally infeasible to ensure quality control on the web-crawled video data, particularly when it comes to establishing clear temporal boundaries and localization between the pre-condition, action process, and the resulting effect. Additionally, eliminating background objects proves to be a challenging task in this context. Lastly, to draw more conclusive insights into the intrinsic evaluations, it would be beneficial to have a human-annotated test set. However, this approach can be prohibitively costly, necessitating further research into semi-automatic methods for conducting such annotations (Huang et al., 2018; Zhang et al., 2022; Dvornik et al., 2023). Regarding our methodology, it is important to note that our focus is on presenting proof-of-concept pretraining tasks rather than achieving optimal performance. Hence, we have not conducted an extensive hyperparameter search, and the results reported in this study are obtained from a single seed run.

Acknowledgments

We are grateful to Matteo Bortoletto for proofreading and providing valuable suggestions for the paper, as well as to Pavel Denisov for offering technical assistance. Lastly, we thank the anonymous reviews for their constructive and valuable feedback.

References

Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald P A Petrick. 2021. Building affordance relations for robotic agents - a review.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. PROST: Physical reasoning about objects through space and time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4597–4608, Online. Association for Computational Linguistics.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. 2023. Affordances from human videos as a versatile representation for robotics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 1–13. IEEE.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020a. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020b. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2017. Simulating action dynamics with neural process networks.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*.

Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. 2015. Mining semantic affordances of visual object categories. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4259–4267.

Nilgun Dağ, Ilkay Atil, Sinan Kalkan, and Erol Şahin. 2010. Learning affordances for categorizing objects and their properties. In *2010 20th International Conference on Pattern Recognition*, pages 3089–3092.

Gautier Dagan, Frank Keller, and Alex Lascarides. 2023. Learning the effects of physical actions in a multi-modal environment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 133–148, Dubrovnik, Croatia. Association for Computational Linguistics.

Atabak Dehban, Lorenzo Jamone, Adam R Kampff, and Jose Santos-Victor. 2016. Denoising auto-encoders for learning of objects and tools affordances in continuous space.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Nikita Dvornik, Isma Hadji, Ran Zhang, Konstantinos G. Derpanis, Animesh Garg, Richard P. Wildes, and Allan D. Jepson. 2023. Stepformer: Self-supervised step discovery and localization in instructional videos. *CoRR*, abs/2304.13265.

Dylan Ebert, Chen Sun, and Ellie Pavlick. 2023. Comparing trajectory and vision modalities for verb representation. *CoRR*, abs/2303.12737.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Ronald A Fisher. 1949. The design of experiments.

Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Chai. 2016. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1814–1824.

Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. 2018. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–945.

James J Gibson. 1977. The theory of affordances. *Hilldale, USA*, 1(2):67–82.

Michael Hanna, Federico Pedeni, Alessandro Suglia, Alberto Testoni, and Raffaella Bernardi. 2022. ACT-thor: A controlled benchmark for embodied action understanding in simulated environments. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5597–5612, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mohammed Hassanin, Salman H. Khan, and Murat Tahtali. 2022. Visual affordance and function understanding: A survey. *ACM Comput. Surv.*, 54(3):47:1–47:35.

He, Zhang, Ren, and Sun. Deep residual learning for image recognition. *and pattern recognition*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

De-An Huang, Shyamal Buch, Lucio M. Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5948–5957. Computer Vision Foundation / IEEE Computer Society.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 1769–1782. PMLR.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J.

Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. 2022. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR.

Esteban Jaramillo-Cabrera, Eduardo F Morales, and Jose Martinez-Carranza. 2019. Enhancing object, action, and effect recognition using probabilistic affordances. *Adapt. Behav.*, 27(5):295–306.

Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: an interactive 3d environment for visual AI. *CoRR*, abs/1712.05474.

Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. 2019a. Mining youtube - A dataset for learning fine-grained action concepts from webly supervised video data. *CoRR*, abs/1906.01012.

Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. 2019b. Mining youtube - A dataset for learning fine-grained action concepts from webly supervised video data. *CoRR*, abs/1906.01012.

Beth Levin and Hovav Malka Rappaport. 2010. Lexicalized scales and verbs of scalar change.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Daniel Loureiro and Alípio Jorge. 2018. Affordance extraction and inference based on semantic role labeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 91–96, Brussels, Belgium. Association for Computational Linguistics.

Liangsheng Lu, Wei Zhai, Hongchen Luo, Yu Kang, and Yang Cao. 2022. Phrase-based affordance detection via cyclic bilateral interaction. *CoRR*, abs/2202.12076.

Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. 2021. One-shot affordance detection. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 895–901. ijcai.org.

Jack Merullo, Dylan Ebert, Carsten Eickhoff, and Ellie Pavlick. 2022. Pretraining on interactions for learning grounded affordance representations. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 258–277, Seattle, Washington. Association for Computational Linguistics.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9876–9886. Computer Vision Foundation / IEEE.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Tushar Nagarajan and Kristen Grauman. 2020. Learning affordance landscapes for interaction exploration in 3d environments. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 1–9. Association for Computational Linguistics.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 564–571.

Michele Persiani and Thomas Hellström. 2019. Unsupervised inference of object affordance from text corpora. In *22nd Nordic Conference on Computational Linguistics (NoDaLiDa'19), September 30 – October 2, 2019, Turku, Finland*. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Erol Şahin, Maya Çakmak, Mehmet R Doğar, Emre Uğur, and Göktürk Üçoluk. 2007. To afford or not to afford: A new formalization of affordances toward Affordance-Based robot control. *Adapt. Behav.*, 15(4):447–472.

Shailaja Keyur Sampat, Pratyay Banerjee, Yezhou Yang, and Chitta Baral. 2022a. Learning action-effect dynamics for hypothetical vision-language reasoning task. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5914–5924, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shailaja Keyur Sampat, Maitreya Patel, Subhasish Das, Yezhou Yang, and Chitta Baral. 2022b. Reasoning about actions over visual and linguistic modalities: A survey. *CoRR*, abs/2207.07568.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15617–15629. IEEE.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. COIN: A large-scale dataset for comprehensive instructional video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1207–1216. Computer Vision Foundation / IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591.

Chao Xu, Yixin Chen, He Wang, Song-Chun Zhu, Yixin Zhu, and Siyuan Huang. 2022. Partafford: Part-level affordance discovery from 3d objects. *CoRR*, abs/2202.13519.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5534–5542. IEEE Computer Society.

Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. 2021. PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2040–2050, Online. Association for Computational Linguistics.

Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. 2022. One-shot object affordance detection in the wild. *Int. J. Comput. Vis.*, 130(10):2472–2500.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13664 of *Lecture Notes in Computer Science*, pages 492–510. Springer.

Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. 2023. Procedure-aware pretraining for instructional video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10727–10738. IEEE.

Yixin Zhu, Yibiao Zhao, and Song-Chun Zhu. 2015. Understanding tools: Task-oriented object modeling, learning and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2855–2864. IEEE Computer Society.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David F. Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3537–3545. Computer Vision Foundation / IEEE.

A Appendix

A.1 Result Verbs: Preprocessing Details

VerbNet. We use VerbNet 3.3 listed in this repository: `https://github.com/cu-clear/verbnet` and the semantic link to find the mapping between VerbNet and FrameNet: `https://github.com/cu-clear/semlink/tree/master/instances`.

imSitu. The list of invalid imSitu roles at the second position we excluded is ["place", "tool", "location", "manner", "instrument", "listener", "container", "model", "suspect", "victimpart", "addressee", "confronted", "start", 'message', 'skill', 'ailment', 'focus', 'resource', 'experiencer', 'phenomenon', 'agentpart', 'coagent', 'end', 'recipient', 'audience', 'blow', 'supported', 'interviewee', 'destination', 'source', 'carrier', 'entityhelped', 'center', 'reciever', 'event', 'naggedperson', 'obstacle', 'stake', 'coparticipant', 'seller', 'performer', 'student', 'giver', 'reference', 'adressee', 'competition', 'occasion', 'image', 'coagentpart', 'bodypart', 'boringthing', 'victim', 'follower', 'perceiver', 'imitation', 'admired', 'chasee', 'undergoer', 'path', 'shelter', 'restrained']. The imSitu dataset (Yatskar et al., 2016) could be downloaded under the link: `https://github.com/my89/imSitu`.

FrameNet. We request version 1.7 from this link: `https://framenet.icsi.berkeley.edu/fndrupal/framenet_request_data`.

Result Verbs: Sure Cases v.s. Unsure Cases. Following the series of automatic steps described in Section 3.1, we derive in total **377** result verb candidates, including **236** sure cases and **141** unsure cases, along with their corresponding **150** unique FrameNet frames. In this work, only the sure cases are utilized for locating the video clips. Some examples of sure cases v.s. unsure cases are shown in Table 7.

A.2 HowTo100M Subtitles: Preprocessing Details

We focus on 13 video categories (the released file, `HowTo100M_v1.csv`, contains 19 categories) that have a denser distribution of result verbs according to a preliminary study conducted on a subset containing 500 videos from each category. In particular, a video category of our interest should have more than 15 unique result verb types, and each of the result verb types within should have more than 100 video clips. The distribution of result verb type across video categories computed during the preliminary study can be seen in Figure 4.

To further downsample the video pool, we select the top 15 viewed videos per wikiHow task id (Miech et al. (2019) search and collect videos based on wikiHow task title. On average, there are 48 videos per task id).

We use the pre-processed version of clip-subtitle, `raw_caption_superclean.json`. All the relevant files can be downloaded from this link: `https://www.di.ens.fr/willow/research/howto100m/`.

The comparison between extracted CAE clips and the non action-centric clip of HowTo100M can be found in Figure 5.

A.3 CAE Split Statistics

Split Details. Within a FrameNet frame, we randomly assign (seed 42) 80%/20% of the verbs (lexical units) to seen/unseen verb classes. As the video frame visual feature extractor is trained with Kinetics400, we ensure no information leakage by controlling the seen verb classes that do not contain Kinetics400 (Kay et al., 2017) verb types. To control the split, we adopt the recommended practices by Xian et al. (2017). For seen verb classes, the corresponding video clips ratio of train-dev-test follows 80%-10%-10% whereas for unseen verb classes, it follows 0%-50%-50%. As for the video

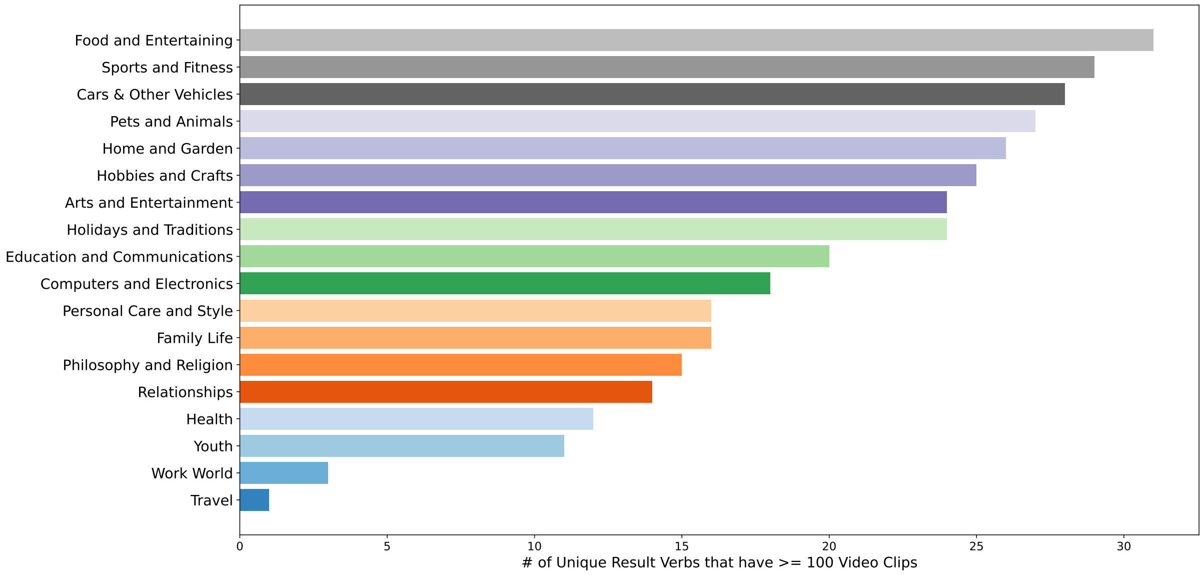


Figure 4: Number of unique result verbs across video categories

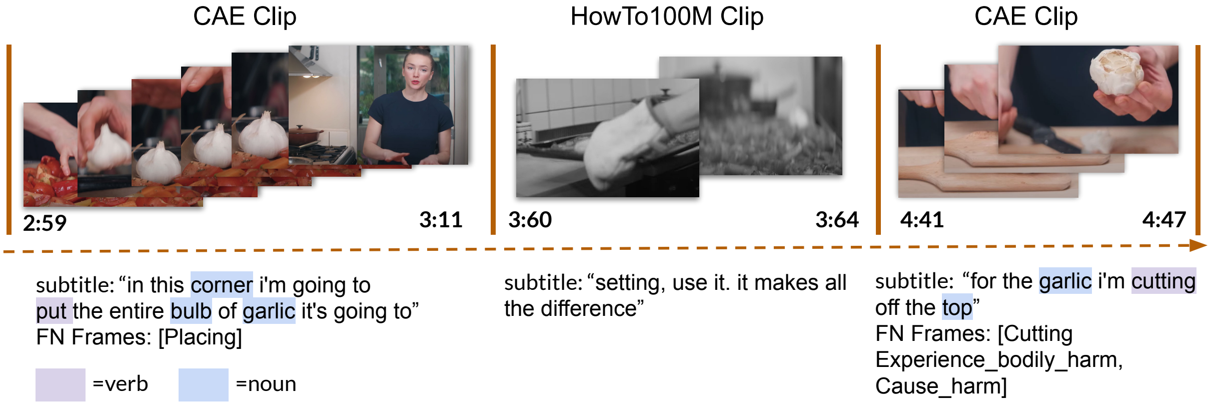

CAE Clip
HowTo100M Clip
CAE Clip
2:59
3:11
3:60
3:64
4:41
4:47
subtitle: "in this corner i'm going to put the entire bulb of garlic it's going to"
FN Frames: [Placing]
=verb
=noun
subtitle: "setting, use it. it makes all the difference"
subtitle: "for the garlic i'm cutting off the top"
FN Frames: [Cutting Experience_bodily_harm, Cause_harm]

Figure 5: Clip–subtitle pairs: CAE vs. HowTo100M, CAE targets action-centric video clips in the HowTo100M

Judgement		Visualness	Effect-causing
Sure Result Verbs	Verb Senses		
attach	[attach-22.3-2-1*, attach*]	✓	✓
bend	[bend-45.2*]	✓	✓
chop	[chop-18.2*, chop-21.2-2*, chop*]	✓	✓
stretch	[stretch-45.2*, stretch*]	✓	✓
tie	[tie-22.4*, tie-22.1-2*, tie^{s}tar]	✓	✓
Unsure Result Verbs	Verb Senses		
activate	[activate-45.4*]	?	✓
block	[block*]	✓	?
carve	[carve-23.3*, carve-21.2-2*, carve*]	✓	?
sniff	[sniff*]	✓	?
warm	[warm-45.4*]	?	✓

Table 7: Result verbs: sure & unsure cases examples. The verb senses are sourced from different resources: * represents VerbNet while $\star$ indicates imSitu (no specific verb sense is recorded). "✓" implies the property is satisfied, whereas "?" indicates uncertainty.

domain, validation and test set are ensured to be similar.

Statistics. For the top 20 and bottom 20 seen verb classes with their top 5 nouns combinations and the corresponding FrameNet frames they can evoke, refer to Tables 8 and 9 respectively. As for the unseen verb classes, refer to Tables 10 and 11. The video clip counts (log scale) of top 100 seen verbs in the train set and unseen verbs (93 classes) in the test set can be found in Figures 6 and 7 respectively.

The co-occurrence heatmap between the top 100 seen verbs and the top 30 nouns in the train set is shown in Figure 8. As for that of unseen verbs (93 classes) and the top 30 nouns in the test set is shown in Figure 9.

A.4 Model Details

Subtitle Input. Subtitles are tokenized with `RobertaTokenizerFast` that uses byte-level Byte-Pair-Encoding from the Hugging Face library (Wolf et al., 2020).

Video Input: Visual Feature Extraction. For each video clip, the temporal context is extended to **3 seconds** before the starting point and after the endpoint of the original time stamp such as to ensure that the pre-condition and the post-condition are sufficiently captured. The visual features are then extracted every 2 seconds (0.5 FPS) with ResNet (He et al.), pretrained on ImageNet (Deng et al., 2009), and with SlowFast (Feichtenhofer et al., 2019), pretrained on Kinectics-400 (Kay et al., 2017) for 2D and 3D features, respectively. Therefore, for a video segment $V = \{\mathbf{v}_i\}_{i=1}^{|V|}$ ($|V|$ is the number of decoded video frames), the resulting visual representation is the concatenation of 2D (2048) and 3D features (2304) ($V \in \mathbb{R}^{|V|} \times 4352$).

HERO Architecture. Designed by Li et al. (2020), it can be viewed as a single-stream vision–language model that captures contextualized video embeddings in a hierarchical fashion. There are three major components: the **Input Embedder**, the **Cross-Modal Transformer**, and the **Temporal Transformer**. The **Input Embedder** comprises the Visual and Textual Embedder. The final textual representation of a sub-word token is obtained by applying a normalization (LN) layer on top of the sum of (1) the token embedding (2) the position embedding and (3) the segment type embedding (one segment type). As for the visual part, the video frame embedding is obtained by projecting the extracted visual features to the same dimension as the token embedding with a fully-connected (FC) layer. Similarly, to obtain the final visual representation, (1) the video frame embedding (2) the position embedding and (3) the segment type embedding comprising three types: `[BEF]` (to denote the pre-condition), `[ACT]` (to denote the action process), `[AFT]` (to denote the post-condition) are summed up and fed through a LN layer. Finally, the multimodal input is the concatenation of the

Top 20 Seen Verbs	Top 5 Nouns	FrameNet frames
make	[video, thing, recipe, time, lot]	[Building]
put	[top, bit, water, side, thing]	[Placing]
cut	[piece, half, knife, top, side]	[Experience_bodily_harm, Cause_harm, Cutting]
cook	[minute, time, cook, water, heat]	[Apply_heat, Cooking_creation, Absorb_heat]
turn	[heat, side, water, minute, light]	[Cause_change, Undergo_change]
pull	[side, loop, top, yarn, thing]	[Cause_motion, Manipulation, Earnings_and_losses]
set	[minute, side, timer, time, top]	[Intentionally_create, Arranging, Change_of_consistency]
stick	[bottom, side, top, hand, pan]	[Cause_motion, Placing, Attaching, Being_attached]
dry	[time, hour, minute, dry, water]	[Cause_to_be_dry]
bake	[soda, powder, minute, oven, teaspoon]	[Cooking_creation]
build	[thing, house, video, time, build]	[Building]
throw	[thing, bit, water, top, stuff]	[Cause_motion]
fold	[half, side, edge, paper, corner]	[Reshaping]
push	[side, button, top, place, way]	[Cause_motion, Manipulation, Cause_change_of_position_on_a_scale]
chop	[onion, garlic, cup, chop, tomato]	[Cause_harm, Cutting]
click	[video, link, button, channel, icon]	[Cause_impact, Impact, Motion_noise]
wash	[hand, water, hair, wash, face]	[Removing, Grooming]
boil	[water, minute, boil, egg, cup]	[Cause_harm, Apply_heat, Absorb_heat]
drop	[drop, water, comment, oil, top]	[Cause_motion]
wrap	[wrap, paper, yarn, wire, tape]	[Placing]

Table 8: Top 20 seen verbs and their top 5 noun combinations

Bottom 20 Seen Verbs	Top 5 Nouns	FrameNet frames
total	[cup, dollar, inch, car, gram]	[Amounting_to, Adding_up]
decay	[matter, decay, wood, time, tree]	[Rotting]
lash	[line, lash, mascara, bottom, thing]	[Attaching]
belt	[belt, pulley, thing, side, drum]	[Cause_harm]
cement	[place, thing, post, cement, piece]	[Attaching]
pulverize	[processor, thing, blender, garlic, salt]	[Grinding]
demolish	[thing, everything, whole, house, button]	[Destroying]
paddle	[paddle, pool, way, river, time]	[Corporal_punishment]
collide	[way, hold, another, fact, time]	[Cause_impact]
plaster	[wall, plaster, giordano, video, time]	[Attaching]
extinguish	[fire, flame, time, water, candle]	[Putting_out_fire]
boat	[boat, water, store, type, hour]	[Operate_vehicle]
devastate	[garden, worm, frost, plant, guy]	[Destroying]
consolidate	[item, power, everything, space, surface]	[Cause_to_amalgamate]
vaporize	[chamomile, water, oil, day, gas]	[Destroying, Change_of_phase]
commute	[work, people, time, day, thing]	[Travel]
shush	[mouth, kill, shush, cup, sauce]	[Silencing]
exterminate	[rat, bedbug, time, garden, pest]	[Killing]
hurl	[time, point, way, wall, wolf]	[Cause_motion]
paw	[ear, paw, hand, food, luke]	[Manipulation]

Table 9: Bottom 20 seen verbs and their top 5 noun combinations

Top 20 Unseen Verbs	Top 5 Nouns	FrameNet frames
mix	[bowl, ingredient, water, everything, mix]	[Cause_to_amalgamate]
place	[top, side, bowl, piece, pan]	[Placing]
break	[piece, egg, thing, time, break]	[Experience_bodily_harm, Cause_harm, Cause_to_fragment, Render_nonfunctional]
roll	[dough, ball, roll, pin, piece]	[Cause_motion, Reshaping, Motion, Mass_motion, Body_movement]
paint	[paint, color, wall, thing, top]	[Filling, Create_physical_artwork, Create_representation]
burn	[burn, bottom, hand, time, fire]	[Experience_bodily_harm, Cause_harm]
attach	[piece, side, top, end, wire]	[Attaching]
blend	[blender, everything, color, water, ingredient]	[Amalgamation, Cause_to_amalgamate]
connect	[wire, side, line, piece, end]	[Attaching]
lift	[side, top, lid, thing, foot]	[Cause_motion]
squeeze	[juice, lemon, water, lime, bit]	[Manipulation]
glue	[piece, glue, place, top, side]	[Building, Attaching]
insert	[hook, hole, toothpick, needle, center]	[Placing]
crush	[garlic, pepper, tomato, ice, clove]	[Cause_harm, Reshaping, Grinding]
roast	[oven, minute, pan, chicken, seed]	[Apply_heat]
rest	[minute, hour, top, time, oven]	[Placing]
hook	[wire, hook, hose, side, thing]	[Attaching]
knock	[door, thing, knock, sock, air]	[Cause_motion]
damage	[plant, hair, root, skin, area]	[Damaging]
split	[half, two, wood, middle, piece]	[Cause_to_fragment]

Table 10: Top 20 unseen verbs and their top 5 noun combinations

Bottom 20 Unseen Verbs	Top 5 Nouns	FrameNet frames
eject	[water, cd, button, air, shell]	[Removing]
shelve	[unit, shelf, side, thing, space]	[Placing]
sliver	[almond, cup, nut, onion, garlic]	[Cause_to_fragment]
pare	[knife, side, skin, line, chisel]	[Cutting]
dissect	[earthworm, poem, dissect, image, way]	[Cause_to_fragment]
rivet	[rivet, handle, place, tang, bottom]	[Attaching]
claw	[way, claw, hand, crab, top]	[Cause_harm]
spear	[spear, fish, hole, stick, lot]	[Cause_harm]
unify	[color, team, look, field, thing]	[Amalgamation, Cause_to_amalgamate]
club	[club, food, another, friend, final]	[Cause_harm]
fracture	[bone, fracture, wall, time, crack]	[Cause_harm, Cause_to_fragment]
thump	[thump, table, video, side, sound]	[Cause_impact, Impact, Make_noise]
obliterate	[time, deck, chop, piece, everything]	[Destroying]
splinter	[wood, splinter, board, side, edge]	[Cause_to_fragment]
jumble	[word, stuff, guy, thing, everything]	[Cause_to_amalgamate]
cleave	[cleave, bit, pinching, wood, log]	[Cause_to_fragment]
manicure	[nail, lawn, hand, manicure, course]	[Grooming]
punt	[puppy, thing, ball, way, punt]	[Cause_motion]
prowl	[prowl, top, treat, attacker, river]	[Self_motion]
mutilate	[beak, man, listener, madhubala, money]	[Cause_harm]

Table 11: Bottom 20 unseen verbs and their top 5 noun combinations

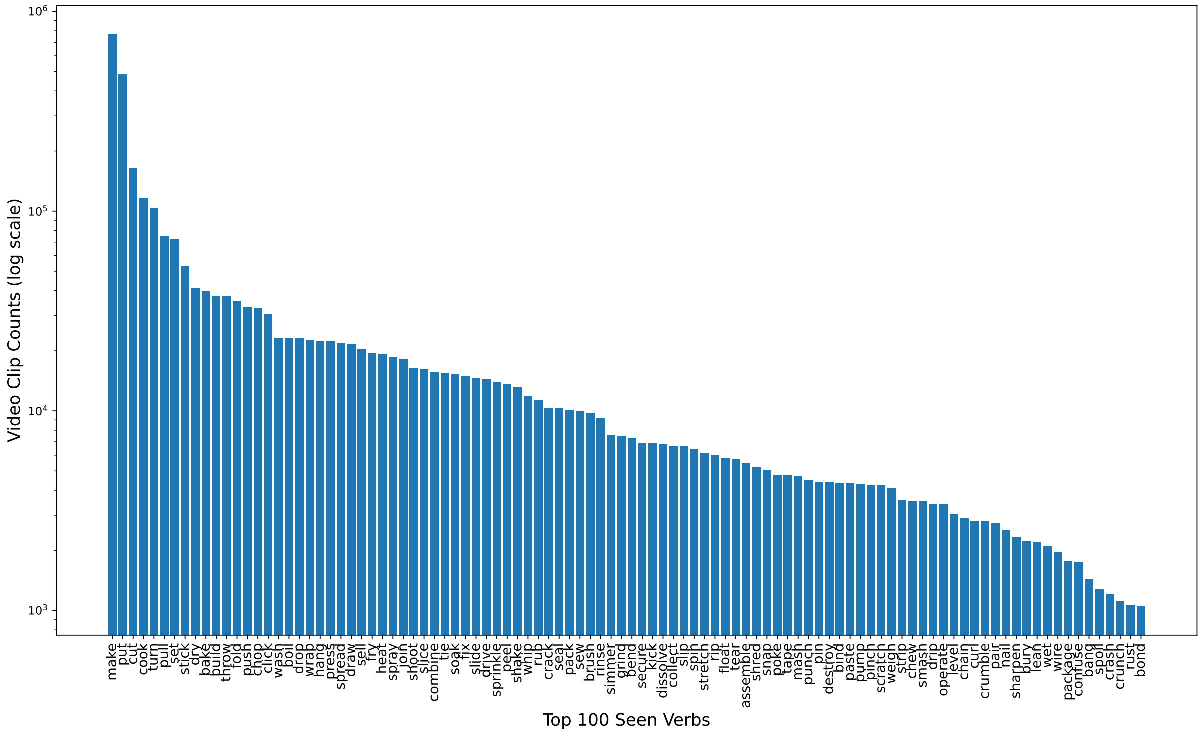


Figure 6: Video clips counts (log scale) of top 100 seen verb classes in the train set

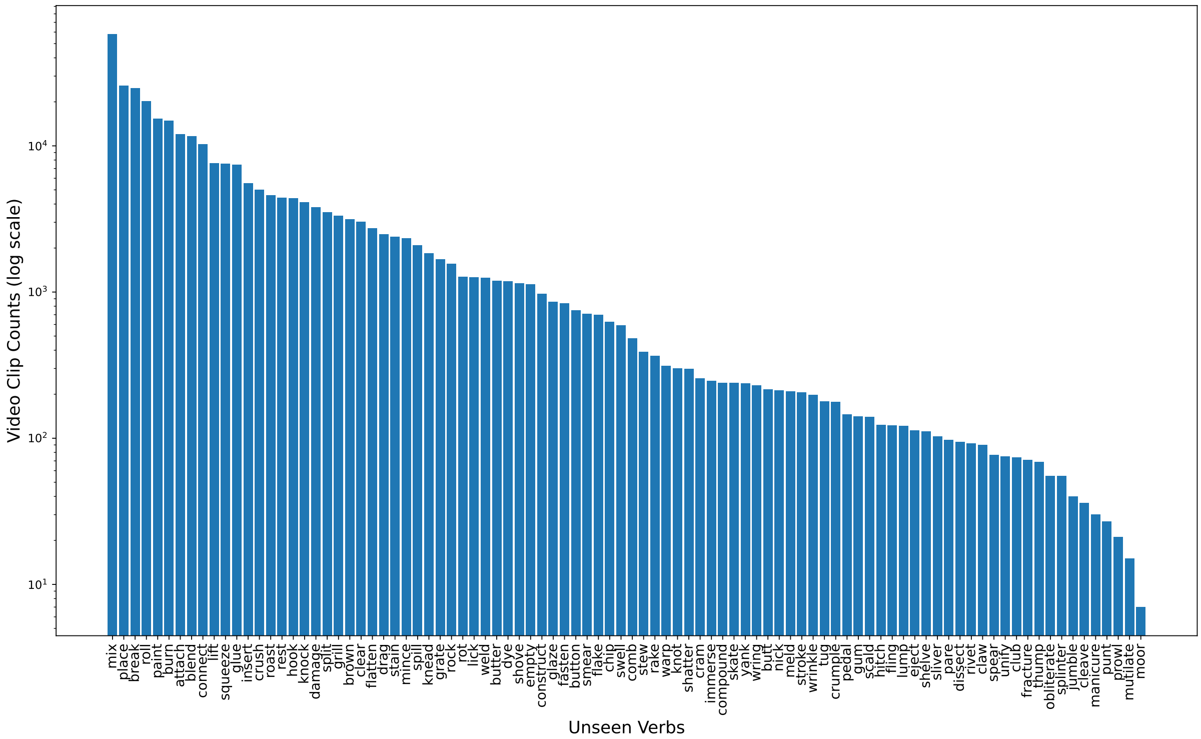


Figure 7: Video clips counts (log scale) of all unseen verb classes in the test set

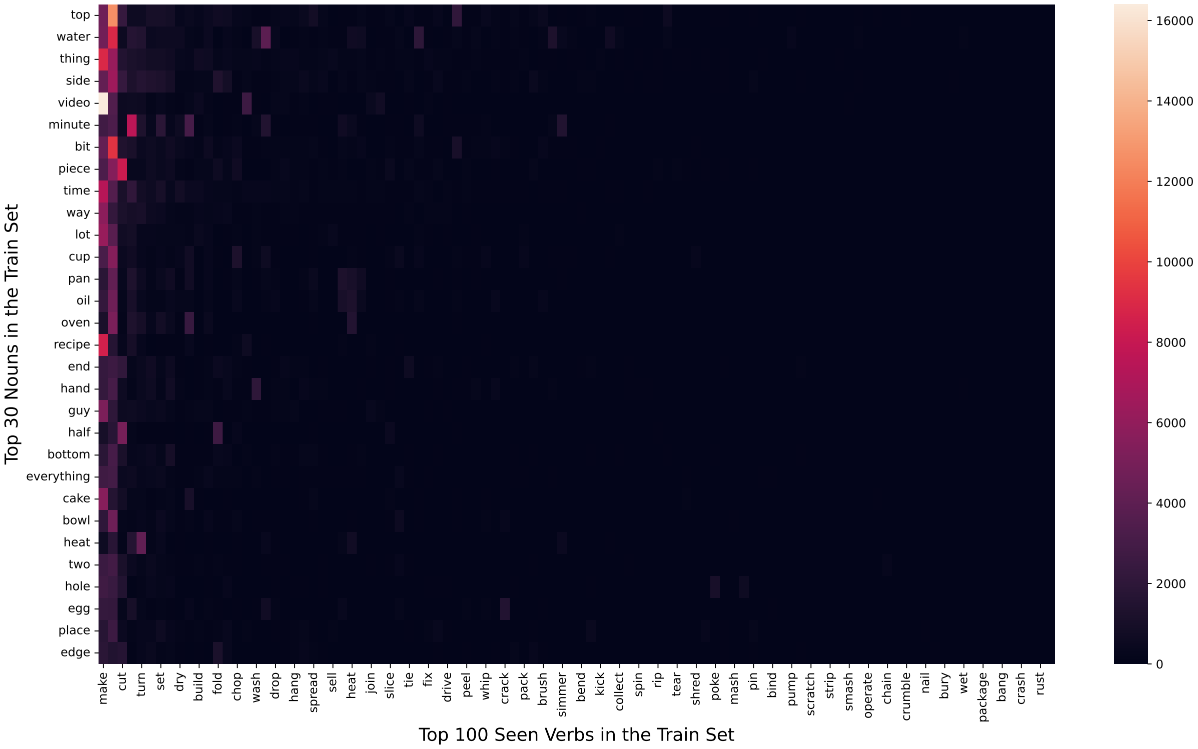


Figure 8: The co-occurrence heatmap between the top 100 seen verb classes and the top 30 nouns in the train set

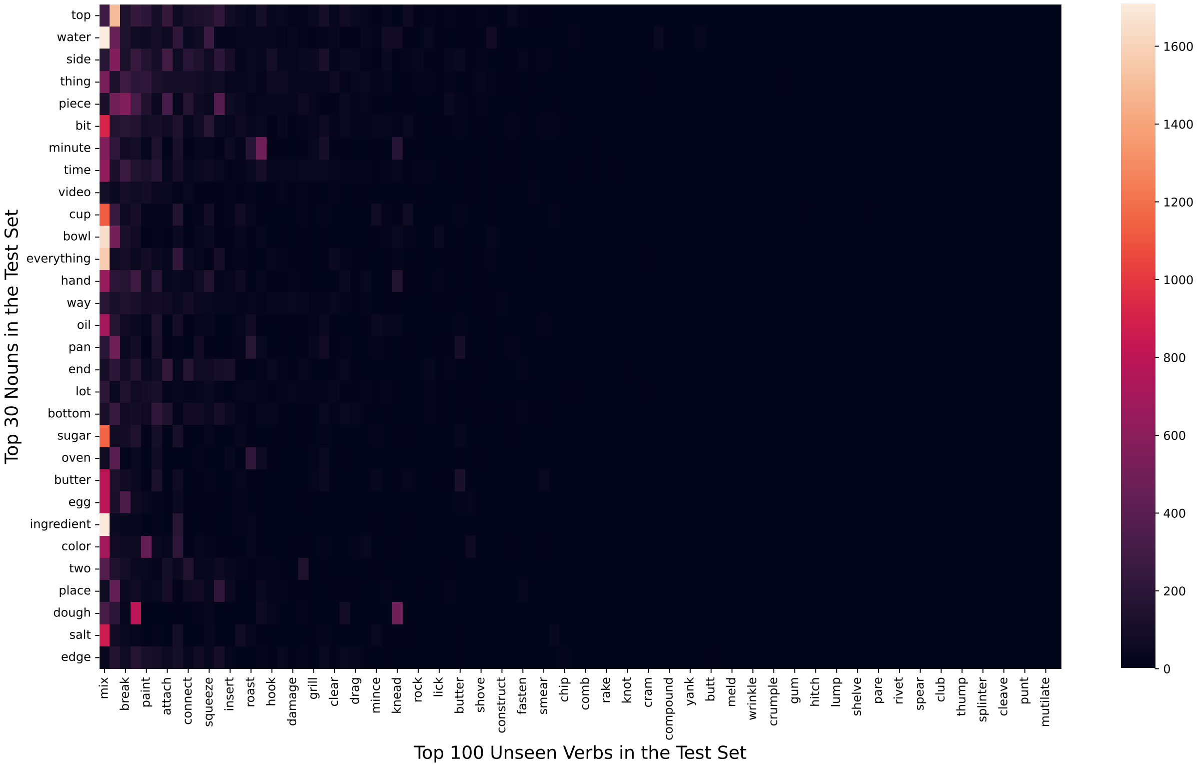


Figure 9: The co-occurrence heatmap between top 100 unseen verb classes and top 30 nouns in the test set

textual and the visual representation.

To learn the implicit association between subtitle tokens and video frames, the **Cross-Modal Transformer** is used to perform cross-modal attention on the multimodal input. The outputs are then contextualized subtitle embeddings and video frame embeddings.

To further obtain temporal-aware video frame embeddings, the local output of video frame embeddings from the Cross-Modal Transformer are fed into the **Temporal Transformer** yielding the global embeddings. The final contextualized video frame embeddings are obtained by adding the local and the global embeddings via a residual connection (He et al., 2016).

A.5 MAM

During the development stage performed on 10% of the training data and test data , we explored three masking strategies (see Table 12):

- **Verb-only**: mask the result verb only.
- **Verb-random-joint**: mask the result verb & and some random words within one data point in the meantime. In this way, more masked tokens have to be reconstructed; thus, it is more challenging.
- **Verb-random-alter**: mask the result verb or random words alternatively, i.e., 50% of the training data with only the result verb tokens being masked, and the rest with only random tokens being masked.

The best action generalization ability (highest accuracy on unseen verbs) was yielded by **verb-random-joint** (see Tab. 13); therefore, we adopt this masking strategy for MAM.

Qualitative Analysis. Figure 10 shows an example of various models' prediction on MAP task. Under limited textual context, MAM-VL$_{Rnd}$ is able to infer the action by reasoning upon the perceptual effect. Figure 11 gives an example of MAM-VL$_{Rnd}$'s extrapolation ability of seen to unseen verbs in terms of visual effect similarities. Performing the *paint* action on the object *canvas*, *spread* on *surface*, and *spray* on *container* result in similar post-conditions, namely, the surface is covered by a layer.

Generalization Analysis. Figure 12 gives two wrong prediction cases of MAM-VL$_{Rnd}$ and MAM-L$_{Rnd}$, where we can see MAM-VL$_{Rnd}$ is able to predict a verb that captures relations with the reference verb from the perspective of situational and lexico-taxonomic semantics.

A.6 MEM

During the development stage performed on 10% of the training data and test data, we explored three masking strategies on controlling the negative video frames. In addition, each candidate video frame has its corresponding video type: [BEF] for the pre-condition, [ACT] for the action process, and [AFT] for the post-condition:

- **Randomized subclips**: randomized [BEF], [ACT] and [AFT] subclips across video clips.
- **Video-based subclips**: [BEF], [ACT] and [AFT] subclips across multiple video clips that belong to the same **video id**.
- **Object-based subclips**: [BEF], [ACT] and [AFT] subclips across multiple video clips that have the same object identified in the video clip.

Details on Controlling the Negative Video Frame Samples. In order to enhance the model's understanding in disentangling the pre-condition, the action process, and the post-condition in the visual space, we consider having the intra-video-clip video frames of types such as [BEF] (pre-condition) and [ACT] (post-condition) in the candidate set of the input. However, we exclude [AFT] video frames within the same clip as it becomes difficult and potentially unfair for the model to distinguish from due to its close temporal proximity.

Quantitative Results Across Different Negative Sampling Strategies During Development. As displayed in Table 14, the test set constructed based on a **video-based** sampling strategy is the most challenging one across three MEM models, despite that it yields the lowest average negative video frames in a batch (avg. neg is 8). In general, when the negative samples in the test set are constructed identically as the train set, the corresponding trained model has the highest accuracy, e.g., the video-based model has +2pp acc over the randomized and object-based one. Moreover, both

MAM Masking Strategies	Original text	Masking Result
verb only	Chop the carrot into pieces. Chop the garlic into small chunks.	[MASK] the carrot into pieces. [MASK] the garlic into small chunks.
verb random joint	Chop the carrot into pieces. Chop the garlic into small chunks.	[MASK] the [MASK] into pieces. [MASK] the [MASK] into small [MASK].
verb random alter	Chop the carrot into pieces. Chop the garlic into small chunks.	[MASK] the carrot into pieces. Chop the garlic into [MASK] chunks.

Table 12: MAM masking strategies

MAM Masking Strategies	Seen Verb Acc	Unseen Verb Acc	Harmonic Mean
Verb-only	25.35	12.93	17.13
Verb-random-joint	28.62	**18.71**	**22.63**
Verb-random-alter	**29.74**	17.09	21.71

Table 13: Intrinsic evaluation of MAM during development on predicting the correct seen / unseen verbs when tested on 10% test set.

video-based and object-based MEM outperform than randomized one when evaluated on the randomized test set, emphasizing the importance of dedicated control on the negative sampling process.

Qualitative Analysis During Development. As seen in Figure 13, the video-based model is able to pick the correct post-condition for *mixing the ingredients* while the randomized-based model struggles to find a correct video frame. We also probe into several failure modes and found some patterns. To begin, the candidate set that contains temporal close clips leads to wrong predictions more often than the temporal distant ones. Concretely, under the wrong predictions set, the average minimum temporal difference within the competitor set is 120 secs, compared to 211 secs for the correct predictions set. Similarly, the candidate set that has the same object occurrences or same action occurrences challenges the model to choose the right [AFT] video frame(s). In the wrong predictions set, 25.66/26.72% of the candidate sets has the same object/action occurrences, compared to 20.98%/18.58% of that for the correct prediction group.

A.7 Pretraining Details

Hyperparameters. Following Li et al. (2020), we used the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $3e-5$, weight decay of 0.01, and warm-up steps of 10000. We ran all the pretraining experiments on 2 NVIDIA RTX A6000 GPUs with a batch size of 16 per single GPU, and gradient accumulation steps (Ott et al., 2018) are set to 2. We set the global training steps to 100000 with maximum training time of 48 hours.

Model Checkpoints. For model variants trained with either MAM or MEM pretraining tasks, the best model checkpoints are saved based on the highest task accuracy on the validation set. Regarding the models that were with the joint task configuration, specifically MULTI-CAE, the final model checkpoint is saved based on the best validation accuracy achieved on both tasks. Note that, for MULTI-CAE-based models, each task undergoes training on only 50% of the entire train set.

Ablated Models. We achieve modality ablation via inter-modal attention masking. More specifically, the weights of MAM-L$_{Rnd}$ are updated with the loss function:

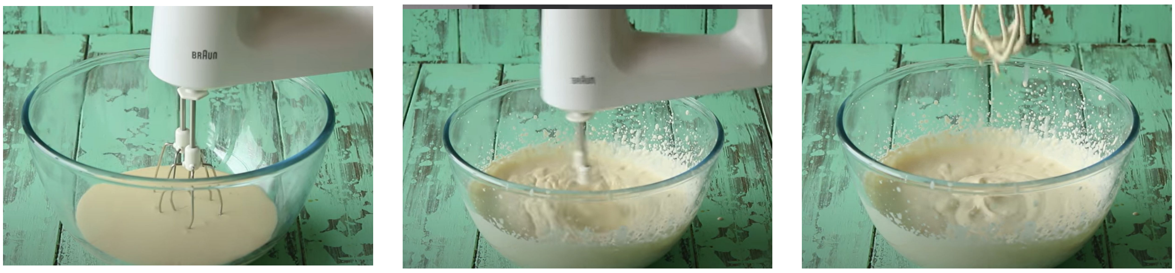
$$\mathcal{L}_{MAM}(\theta) = -\log P_\theta(\mathbf{s}_{a^*}|S_{\backslash \mathbf{s}_{a^*}})$$

As for MEM-V, we update its weights with the loss function:

$$\mathcal{L}_{MEM}(\theta) = -\log P_\theta(\mathbf{v}_{\texttt{[AFT]}}|V_{\backslash \mathbf{v}_{\texttt{[AFT]}}})$$

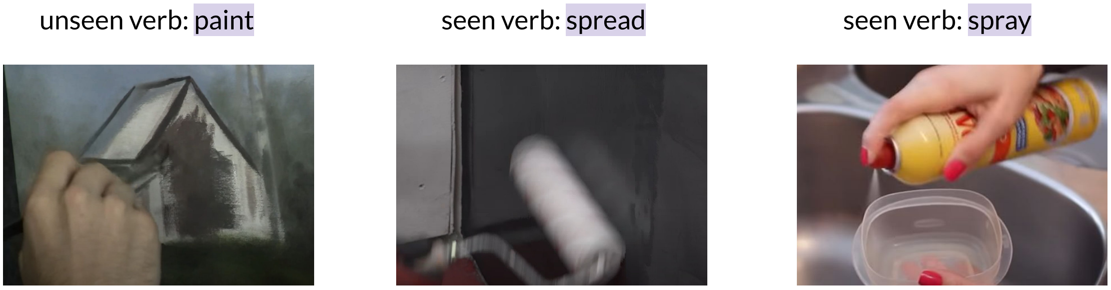
A.8 PROST Task

Zero-shot Inference. The task is formulated as a cloze style task, where the [MASK] token should be reconstructed with an object in a set of 4 mentioned candidates in the context that can afford the action. Following Aroca-Ouellette et al. (2021),



subtitle: "and [MASK] that up until it reaches stiff"
reference verb: whip
(correct) MAM-VL-RND: whip
(wrong) MAM-VL-RND (on L): combine
(wrong) MAM-L-RND (on L): make

Figure 10: Action inference task: MAM-VL$_{Rnd}$ is able to infer the correct action "whip" by reasoning upon the visual causal effect process. Under the language-only inference setting, models fail as the subtitle gives limited information.



unseen verb: paint

seen verb: spread

seen verb: spray

subtitle: "paint in as well as you can/ what i mean is really working into the canvas/it's"

subtitle: "ideal for smooth surfaces/ spread it from bottom to top/ in vertical strips until"

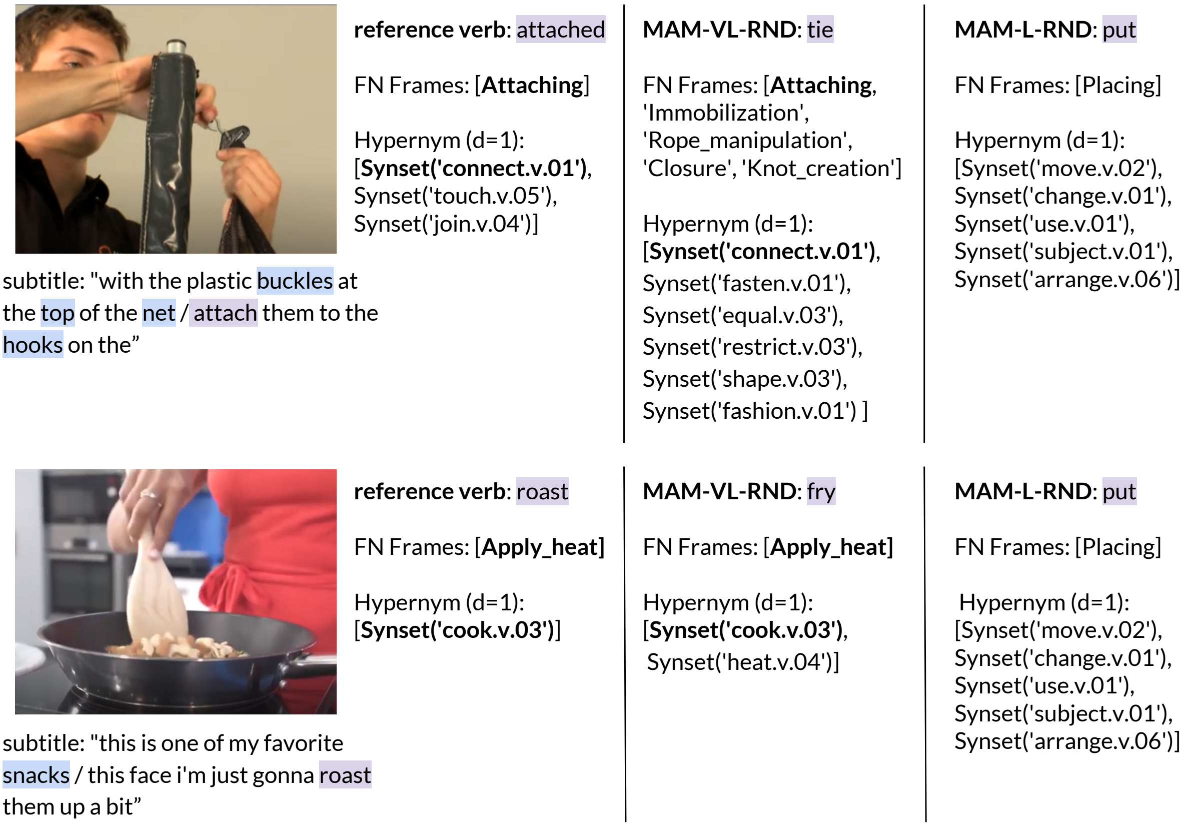
subtitle: "spray some in the container before you"

Figure 11: Action generalization: MAM-VL$_{Rnd}$ generalizes to the unseen verb "paint". Its causal effect is similar to the two seen verbs "spread" & "spray" that produce a light coat on the surface of an object. For readability, we use "/" to indicate the sentence boundary.

we obtain the logits of the 4 object candidates instead of the whole vocabulary from the pretrained MAM head. Subsequently, we compute the probabilities of these candidates using the Softmax function. The object with the highest probability is the model's decision.

PROST Probed Object Occurrences in CAE. As shown in Table. 15, all the objects probed in the PROST affordance groups are seen to our models. Note that the statistic does not describe the occurrences of action-object combinations as only "Slide" is a seen verb to our models. In terms of action–object combinations, only [**slide**, *oil*] and [**slide**, *grease*], are seen to our model. We hypothesize that MULTI-CAE-VL and MAM-VL may have leveraged the shared visual properties of these objects, i.e., slipperiness, to extend the affordance understanding on the other objects like *soap* and *frost*.

Model Robustness: Original vs. Inverses. The individual results on original and inverse templates across the concepts are shown in Table 16.



reference verb: attached

FN Frames: [**Attaching**]

Hypernym (d=1):
[**Synset('connect.v.01')**,
Synset('touch.v.05'),
Synset('join.v.04')]

subtitle: "with the plastic buckles at the top of the net / attach them to the hooks on the"

MAM-VL-RND: tie

FN Frames: [**Attaching**,
'Immobilization',
'Rope_manipulation',
'Closure', 'Knot_creation']

Hypernym (d=1):
[**Synset('connect.v.01')**,
Synset('fasten.v.01'),
Synset('equal.v.03'),
Synset('restrict.v.03'),
Synset('shape.v.03'),
Synset('fashion.v.01')]

MAM-L-RND: put

FN Frames: [Placing]

Hypernym (d=1):
[Synset('move.v.02'),
Synset('change.v.01'),
Synset('use.v.01'),
Synset('subject.v.01'),
Synset('arrange.v.06')]

reference verb: roast

FN Frames: [**Apply_heat**]

Hypernym (d=1):
[**Synset('cook.v.03')**]

subtitle: "this is one of my favorite snacks / this face i'm just gonna roast them up a bit"

MAM-VL-RND: fry

FN Frames: [**Apply_heat**]

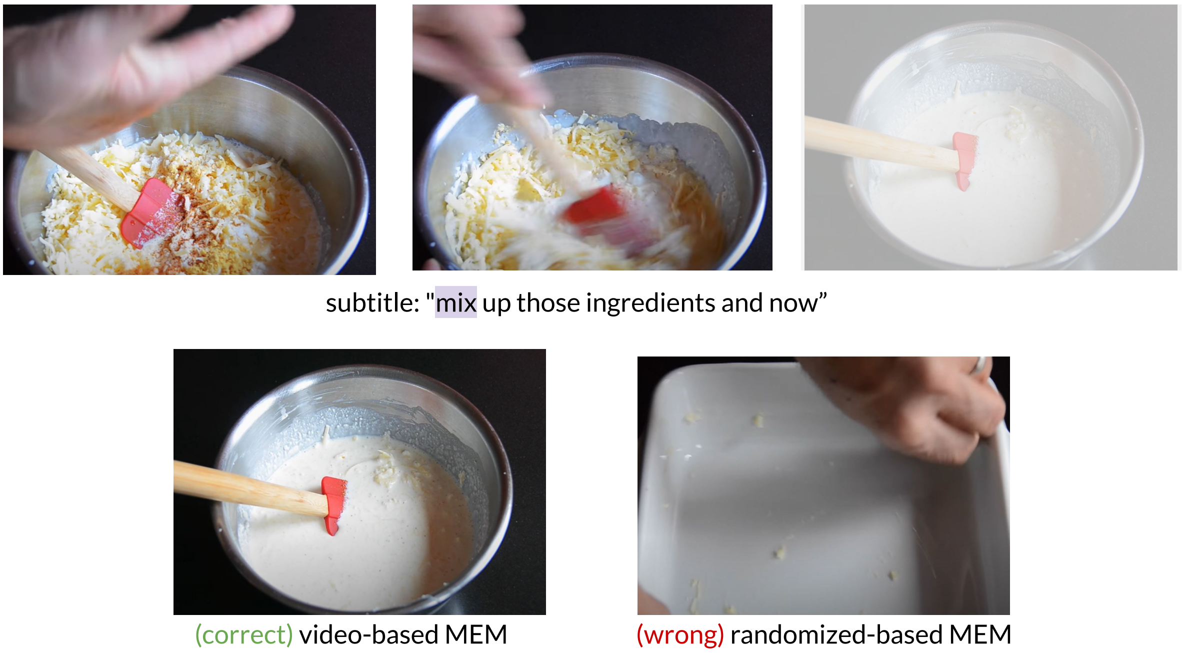
Hypernym (d=1):
[**Synset('cook.v.03')**,
Synset('heat.v.04')]

MAM-L-RND: put

FN Frames: [Placing]

Hypernym (d=1):
[Synset('move.v.02'),
Synset('change.v.01'),
Synset('use.v.01'),
Synset('subject.v.01'),
Synset('arrange.v.06')]

Figure 12: Action generalization: both MAM-VL$_{Rnd}$ and MAM-L$_{Rnd}$ predict the wrong verb, but MAM-VL$_{Rnd}$ predicts the verbs that share the same FrameNet frame and Hypernym (depth=1) with the reference verbs. MAM-L$_{Rnd}$ tends to predict one of the most common seen verbs, "put".

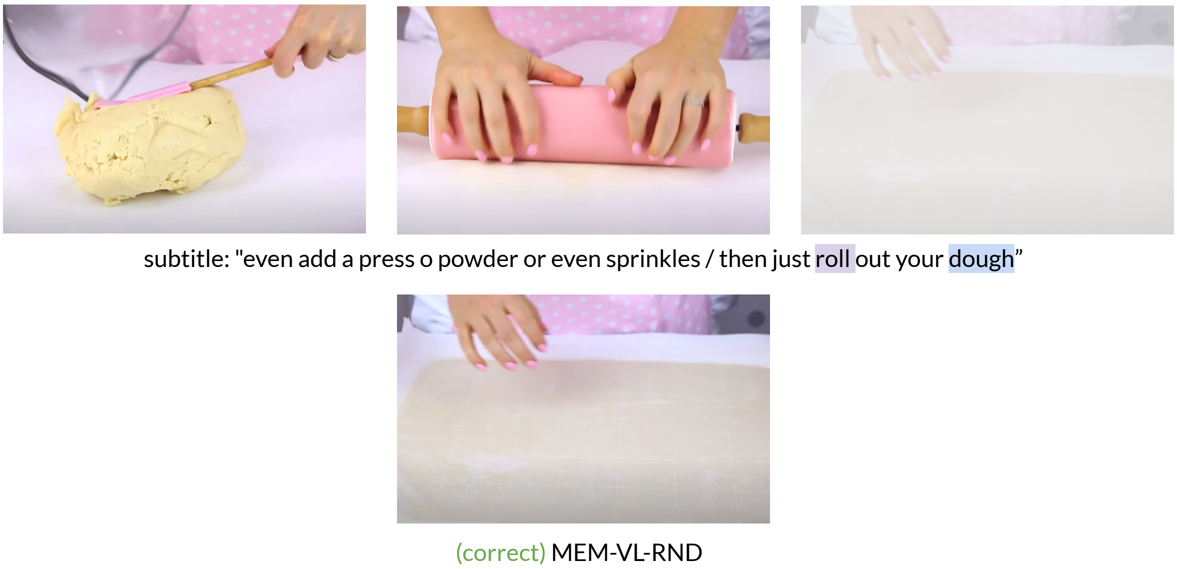


subtitle: "mix up those ingredients and now"

(correct) video-based MEM

(wrong) randomized-based MEM

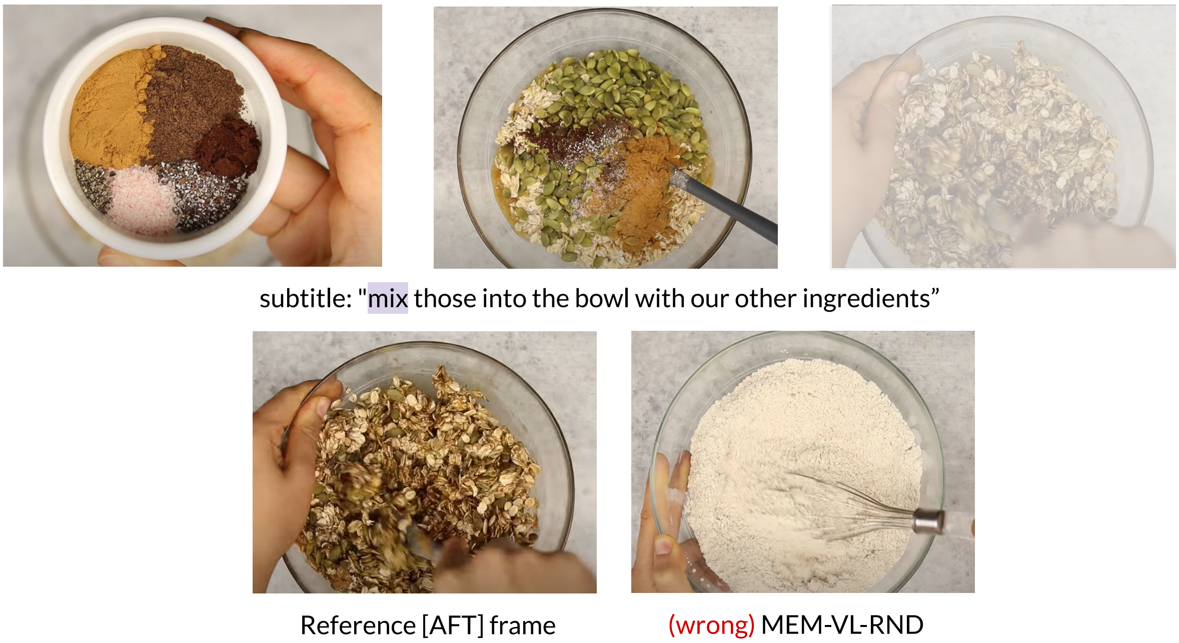
Figure 13: Video-based MEM chose the correct `[AFT]` frame, the randomized one fails.



subtitle: "even add a press o powder or even sprinkles / then just roll out your dough"

(correct) MEM-VL-RND

Figure 14: Entity equivalence: although the action "roll" is unseen to MEM-VL$_{Rnd}$, it has successfully learned that "dough" could be rolled given its "foldability" which is a property shared among many seen objects like "cloth".



subtitle: "mix those into the bowl with our other ingredients"

Reference [AFT] frame

(wrong) MEM-VL-RND

Figure 15: Error analysis: MEM-VL$_{Rnd}$ fails for a sensible reason, it selected the `[AFT]` frame that conceptually follows the instruction process of "mixing ingredients".

MEM Negative Sampling Strategies	Randomized-set (avg.neg=618) Acc	Video-based-set (avg.neg=8) Acc	Object-based-set (avg.neg=452) Acc
Randomized-based	87.74	70.4	87.91
Video-based	88.08	**72.45**	88.04
Object-based	**88.57**	70.82	**89.55**

Table 14: Intrinsic evaluation of MEM during development on predicting the correct video `[AFT]` frame(s) among negatives constructed using various sampling schemes when tested on 10% test set.

Affordance Concepts	Original	Inversed
Stack	book: 481, block: 697, box: 1,632, coin: 113, plate: 1353	ball: 2,040, bottle: 918, egg: 3,083, flower: 1,068, lamp: 108
Roll	apple: 690, ball: 2,040, bottle: 918, egg: 3,083, can: 522	book: 481, block: 697, box: 1,632, mirror: 168, microwave: 517
Grasp	ball: 2,040, block: 697, book: 481, bottle: 918, flower: 1,068	flour: 2,692, rice: 1,233, salt: 2,862, snow: 151, sugar: 3227
Break	bottle: 918, egg: 3,083, glass: 1,254, mirror: 168, plate: 1,353	ball: 2,040, coin: 113, pen: 252, pillow: 222, shirt: 373
Slide	ice: 880, frost: 56, grease: 215, oil: 3,711, soap: 454	carpet: 312, concrete: 246, grass: 229, gravel: 71, rubber: 133
Bounce	asphalt: 10, brick: 189, concrete: 246, rubber: 133, steel: 176	carpet: 312, foam: 290, grass: 229, leave: 1012, snow: 151

Table 15: Affordance concepts probed in the PROST and their corresponding object set, along with the total number of training instances per object in the CAE dataset.

Model	Stack		Roll		Grasp		Break		Slide		Bounce	
	Ori.	Inv.	Ori.	Inv.	Ori.	Inv.	Ori.	Inv.	Ori.	Inv.	Ori.	Inv.
MAM-L	0.0	60.0	30.0	10.0	8.0	40.0	34.0	20.0	44.0	4.0	6.0	40.0
MAM-VL	20.0	24.0	10.0	40.0	2.0	64.0	32.0	20.0	40.0	20.0	20.0	26.0
MULTI-CAE-VL	0.0	68.0	40.0	8.0	2.0	58.0	32.0	20.0	80.0	0.0	0.0	76.0
FLAVA	21.2	16.7	49.5	4.3	0.2	68.8	45.5	3.6	43.6	4.0	8.2	39.9
RoBERTa-B	27.33	27.42	32.17	17.67	20.25	25.42	56.92	5.0	17.83	35.33	26.75	24.33

Table 16: Accuracy of original vs. inverse templates across affordance groups.