Towards Robust Evaluation of Visual Activity Recognition: Resolving Verb Ambiguity with Sense Clustering

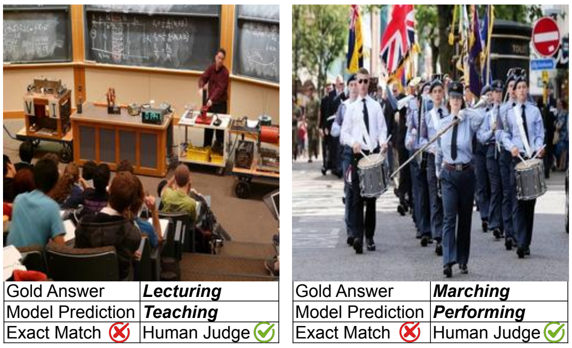
Louie Hong Yao[1] **Nicholas Jarvis**[2] **Tianyu Jiang**[2]
[1]Independent Researcher [2]University of Cincinnati
lhyao731@gmail.com, jarvisns@mail.uc.edu, tianyu.jiang@uc.edu

Abstract

Evaluating visual activity recognition systems is challenging due to inherent ambiguities in verb semantics and image interpretation. When describing actions in images, synonymous verbs can refer to the same event (e.g., *brushing* vs. *grooming*), while different perspectives can lead to equally valid but distinct verb choices (e.g., *piloting* vs. *operating*). Standard exact-match evaluation, which relies on a single gold answer, fails to capture these ambiguities, resulting in an incomplete assessment of model performance. To address this, we propose a vision-language clustering framework that constructs **verb sense clusters**, providing a more robust evaluation. Our analysis of the imSitu dataset shows that each image maps to around four sense clusters, with each cluster representing a distinct perspective of the image. We evaluate multiple activity recognition models and compare our cluster-based evaluation with standard evaluation methods. Additionally, our human alignment analysis suggests that the cluster-based evaluation better aligns with human judgments, offering a more nuanced assessment of model performance.

1 Introduction

Visual activity recognition systems aim to interpret events in images by predicting the primary activity, such as *riding* and *cooking*. However, evaluating these systems presents significant challenges due to multiple levels of ambiguities. One ambiguity comes from the complexity of word meanings and semantic relations. For instance, synonymous verbs can describe the same event: *teaching* and *lecturing* often refer to the same action. Another ambiguity lies within the images. When describing the event in one image from different perspectives, the choice of verbs can vary. For example, an image of a marching band can be described by both *marching* and *performing*. As shown in Figure 1, a model should receive credit for predicting *teaching/performing* despite that the ground truth label is *lecturing/marching*.



Gold Answer	***Lecturing***	Gold Answer	***Marching***
Model Prediction	***Teaching***	Model Prediction	***Performing***
Exact Match ☒	Human Judge ✓	Exact Match ☒	Human Judge ✓

Figure 1: Accuracy score based on the exact match against gold labeling does not consider the ambiguity.

Previous work on visual action recognition has been adopting discrete semantic classes to describe actions, both in static images (Ronchi and Perona, 2015; Chao et al., 2015) and videos (Gu et al., 2018; Wang et al., 2024). Yatskar et al. (2016) introduced the *imSitu* dataset, which has become a widely adopted benchmark for visual activity recognition. The task is defined as, given an image, the model should first identify the appropriate verb describing the image, then fill in the other participants in the event. The imSitu dataset comprises 126,102 images annotated with 504 verbs, covering a broad spectrum of daily activities. Previous approaches to this task (Cooray et al., 2020; Li et al., 2022; Jiang and Riloff, 2023) have been following the standard evaluation based on exact match: a prediction is correct only when the output verb is the same as the gold answer. We argue that this evaluation metric does not reflect the true capability of a recognition system due to two types of ambiguity: (1) **Synonymy** - multiple synonymous verbs that describe the same activity, and (2) **Multi-perspectives** - multiple valid verbs that describe the activity from different perspectives.

To address the ambiguity in evaluation, we pro-

Findings of the Association for Computational Linguistics: EACL 2026, pages 656–672
March 24-29, 2026 ©2026 Association for Computational Linguistics

pose a novel vision-language clustering framework that constructs verb sense clusters. These clusters serve two purposes: (1) quantifying inherent ambiguity in exact-match evaluation and (2) enabling more accurate assessment of visual activity recognition systems. Using both internal and external metrics, we validate the effectiveness of our clustering approach. Based on our sense clusters, we reveal that on average, an image can typically be described using four sense clusters, and each cluster contains close to two synonymous verbs. Through extensive experimentation comparing multiple supervised models and multimodal large language models, we demonstrate that our cluster-based evaluation methodology offers a more comprehensive and accurate assessment than traditional exact-match approaches. Our code is publicly available.[1] In summary, our contributions are three-fold:

1. We propose a novel vision-language clustering framework to build sense groups based on images and activity verbs, and evaluate the clustering quality with both internal and external metrics.
2. We demonstrate that our cluster-based evaluation offers a more accurate and nuanced assessment than exact-match accuracy.
3. We quantify two types of ambiguity in evaluation based on exact match using the imSitu dataset: synonymous verbs describing the same activity and multiple valid verbs capturing different perspectives.

While our experiments focus on imSitu, the proposed clustering-based evaluation framework is designed to be readily adaptable to other visual activity datasets.

2 Related Work

Visual activity recognition has seen significant development across both image and video domains, with datasets of varying granularity and scope. Early examples of image-based datasets include Stanford-40 (Yao et al., 2011), which covers 40 common human activities, and its extension to 89 activities by Le et al. (2013). Subsequent datasets such as COCO-a (Ronchi and Perona, 2015) and HICO (Chao et al., 2015) focus on everyday actions with 140 and 117 classes respectively, while more comprehensive collections like imSitu (Yatskar et al., 2016) feature 504 verbs drawn from FrameNet, and SVO-Probes (Hendricks and Nematzadeh, 2021) cover 421 verbs. Video datasets show similar diversity in their treatment of actions, from AVA's 80 atomic visual actions (Gu et al., 2018) to VidSitu's (Sadhu et al., 2021) extensive coverage of 1,500 verbs from PropBank. More recently, HARDVS (Wang et al., 2024) introduced 300 fine-grained activity categories captured using dynamic vision sensors. This broad spectrum of activity classes across datasets highlights both the complexity of action recognition and the potential utility of our cluster-based evaluation framework.

The imSitu dataset has served as a key benchmark for advancing visual activity recognition techniques. Early approaches employed CNN-based backbones for image encoding (Yatskar et al., 2016; Pratt et al., 2020). More recent work has leveraged CLIP-based models (Li et al., 2022; Roy et al., 2024) and transformer-based models (Cho et al., 2022; Jiang and Riloff, 2023). While these approaches have progressively improved performance, their reliance on exact verb matching for evaluation may underestimate their true capabilities. Although CLIPScore (Hessel et al., 2021) has gained popularity as a flexible evaluation metric for image captioning using CLIP embeddings, its generic similarity measures are not suited for the specific challenges of verb classification. Our work also relates to recent efforts in visual word sense disambiguation (Raganato et al., 2023), which addresses word polysemy in visual contexts.

Semantic evaluation of vision-language models has traditionally relied on metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015), which emphasize n-gram overlap but often miss deeper semantic alignment. More recent metrics such as CLIPScore (Hessel et al., 2021), RefCLIPScore (Hessel et al., 2021), and UMIC (Lee et al., 2021) leverage vision-language embeddings to better assess cross-modal relevance. However, these metrics remain limited in capturing fine-grained semantic distinctions, such as verb sense ambiguity and shifts in perspective.

3 Methodology

We propose a two-step clustering framework that clusters images into verb sense groups. We hypothesize that each cluster represents a fine-grained

[1] https://github.com/ruyi101/multimodal-verb-senses

group of activities that are similar. For example, a cluster of different images depicting a teacher teaching in the classroom, with verbs *lecturing*, *instructing*, *teaching*, *educating* all describing this scenario. We use a multimodal large language model to generate image-verb pairs as nodes and cluster them first by shared verbs, then into fine-grained sense groups. Figure 2 shows our two-step clustering framework.

3.1 Acquiring Image-Verb Pairs

The original imSitu dataset collected the images by querying Google image search with verbs and using those verbs as gold labels for the returned images, which potentially introduces labeling bias. For example, searching for "*drawing*" images might return both picture (a) and (b) in Figure 2. However, when considering the main activity in picture (b), "*drawing*" and "*teaching*" are equally valid answers. To mitigate this bias, we query two multimodal large language models to generate all appropriate verbs to describe each image: GPT-4o mini (Hurst et al., 2024) and Llama-3.2-90B (Dubey et al., 2024). Prompting details can be found in Appendix A.

For each image, we collect a set of verbs through prompting and include the original gold-label verb if it is not already present. We then filter these verbs to retain only those that appear in the imSitu dataset. Next, we construct nodes, where each node represents an `<image,verb>` pair. To capture the joint semantic meaning of the image and verb, we use a multimodal Llama model to generate embeddings by processing both inputs simultaneously. Specifically, we extract the final hidden state after the text-image cross-attention layer and use it as the representation vector for each node.

3.2 Two-Step Clustering

The two-step clustering framework aims to group `<image,verb>` pairs into coherent clusters, so that each cluster represents one type of activities with similar images and verbs. This subsection introduces the procedure, with the corresponding pseudo-code provided in Algorithm 1 in the appendix.

Step 1: Same-Verb Clustering. In the first step, the goal is to disambiguate the fine-grained senses of each individual verb. For each verb v in the list of 504 target verbs $\mathcal{V}$, we first collect all images $\mathcal{I}_v$ associated with v from the dataset $\mathcal{D}$. To capture both the visual and semantic properties of the `<image,verb>` pairs, the images and verbs are transformed into high-dimensional embeddings $\mathcal{E}_v$ using Llama-3.2-11B. Specifically, the verb v is provided as input text, while the corresponding image is provided as the input image. The embeddings are extracted by leveraging the model's cross-attention mechanism, which integrates the semantic information from the verb and the visual information from the image. The final embedding for each `<image,verb>` pair is obtained from the last token of the model's final layer, which is a 4096-dimensional representation.

To cluster the embeddings $\mathcal{E}_v$, we apply two clustering algorithms, K-Means and Hierarchical Agglomerative Clustering (HAC), collectively referred to as $\mathcal{A}$. For K-Means, we normalize the embeddings to unit vectors so that Euclidean distance is monotonically related to cosine distance. When using HAC, we employ complete linkage and cosine distance to group the embeddings. The optimal number of clusters k for each verb is determined by exploring a candidate range of cluster numbers $\mathcal{K} = 2, 3, \ldots, 16$. For each $k \in \mathcal{K}$, the algorithm $\mathcal{A}$ generates a clustering result $\mathcal{C}_k$, which is evaluated using the Silhouette score. The k that achieves the highest Silhouette score is selected as the optimal number of clusters for v, and the corresponding clustering result $\mathcal{C}_k$ is stored.

Step 2: Cross-Verb Clustering. The second step addresses cross-verb ambiguities by clustering the results from Step 1 into final clusters. The clustering algorithm $\mathcal{A}$ used in this step is the same as in Step 1. For example, if K-Means is used for clustering in Step 1, K-Means will also be used in Step 2. These final clusters represent shared meanings across verbs and provide a higher-level understanding of the actions depicted in the images.

Each cluster $c \in \mathcal{C}_{\text{step1}}$ is represented by its average embedding $\mathbf{e}_c$, which captures the semantic and visual properties of the cluster. The number of final clusters is determined by computing $k_r = \text{int}(504 \times r)$, where 504 is the total number of target verbs and r is a ratio selected from the candidate list $\mathcal{R} = \{0.6, 0.7, \ldots, 1.6\}$ with an increment of 0.1. The clustering algorithm $\mathcal{A}$ is then applied to the set of average embeddings $\{\mathbf{e}_c\}$ using k_r clusters, producing the final clustering result $\mathcal{C}_{\text{final}}$. The optimal hyperparameter r, and consequently k_r, is selected based on the Silhouette score. The clustering result $\mathcal{C}_{\text{final}}$ with the highest Silhouette score is chosen as the final output.

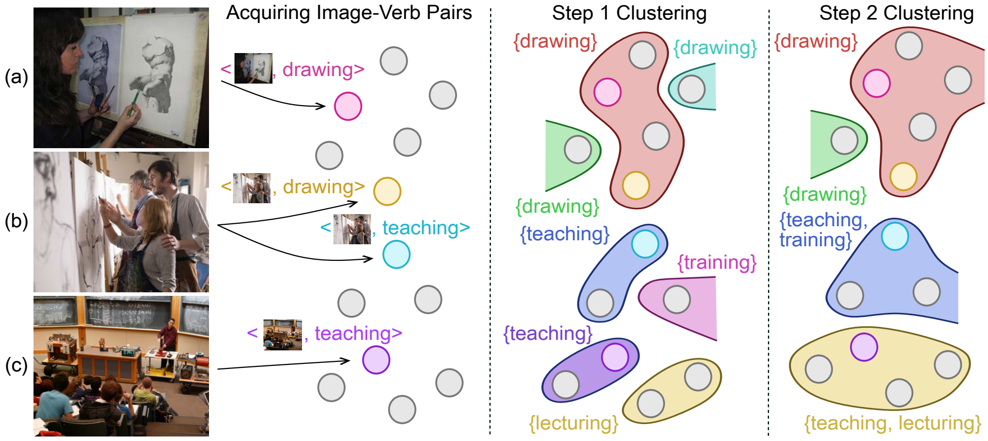


Figure 2: Overview of our two-step clustering framework. Step 1 groups <image, verb> pairs with the same verb into fine-grained sense clusters. Step 2 merges these across verbs to capture shared semantic meanings.

4 Evaluation and Analysis

We organize our analysis into four parts. We begin by evaluating the quality of our clustering framework and use the resulting clusters to analyze sources of ambiguity in the imSitu dataset. Next, we apply these clusters to evaluate a range of visual activity recognition systems. We further examine the robustness of our evaluation method and its alignment with human judgment. In the third part, we provide a qualitative analysis of the learned clusters. Finally, we assess the limitations of CLIP-Score as an evaluation metric for verb-centric activity recognition.

4.1 Quality of Clusterings and Ambiguity Analysis

We begin by evaluating the quality of our clustering results, and then leverage them to analyze sources of ambiguity in the imSitu dataset.

Assessing the Quality of Clusters. To assess the quality of our clustering, we use two internal metrics and one external diagnostic. As internal metrics, we report the Silhouette score (Rousseeuw, 1987) and the Calinski–Harabasz index (Caliński and Harabasz, 1974), which quantify cluster separation and cohesion. As an external reference, we compute a purity score with respect to WordNet synsets (Miller, 1995). For each cluster, we identify all synsets containing at least one verb in the cluster and compute purity using the maximum synset overlap (details in Appendix B).

Importantly, while WordNet synsets encode *linguistic senses* defined by lexical-semantic distinctions, our goal is to capture *visually grounded senses* that reflect how actions are manifested and perceived in images. For instance, verbs such as *running*, *racing*, and *chasing* correspond to distinct WordNet senses, yet they often share highly similar visual configurations involving motion, posture, and spatial relations. Accordingly, WordNet purity contributes to our assessment of clustering quality alongside internal metrics, serving as an auxiliary external diagnostic to detect over-merging of clearly unrelated verb meanings, rather than as a standalone measure of cluster quality.

Clustering Approach	Sil.	C-H	PuS
Llama-3.2-90B + K-Means	**0.22**	6.53	**0.50**
Llama-3.2-90B + HAC	0.18	5.33	0.40
GPT-4o mini + K-Means	**0.22**	**8.05**	0.45
GPT-4o mini + HAC	0.20	6.15	0.42

Table 1: Clustering quality (Sil. = Silhouette, C-H = Calinski-Harabasz, PuS = Purity Score).

In Table 1, we compare clustering results obtained by prompting GPT-4o mini and Llama-3.2-90B, using K-Means and hierarchical agglomerative clustering (HAC). K-Means consistently outperforms HAC across internal metrics, indicating stronger cluster cohesion and separation. While GPT-4o mini achieves higher Calinski–Harabasz scores under K-Means, Llama-3.2-90B yields higher WordNet purity. Although purity is not expected to be perfect, it provides a useful external reference when considered alongside internal metrics. Balancing internal cohesion with this external validation signal, we select the K-Means clustering derived from Llama-3.2-90B for evaluating visual activity recognition systems in Section 4.2.

Ambiguity Analysis. Our clusters can also be used

Metric	GPT-4o mini	Llama-3.2-90B
# of Clusters	554	655
Verbs per Cluster	1.97	1.59
Clusters per Image	4.44	3.61
Multi-Image Rate	0.95	0.71
Clusters per Verb	2.17	2.07
Multi-Verb Rate	0.54	0.54

Table 2: Clustering metrics illustrating ambiguity.

to analyze ambiguities within the imSitu dataset. We use the clusters derived from both GPT-4o mini and Llama-3.2-90B with K-Means to quantify ambiguity in verb usage. Table 2 presents metrics that capture two types of ambiguity:

- *Synonymy.* It is captured when different verbs are grouped into the same cluster. Row 2 (Verbs per Cluster) shows that each cluster contains 1.59 (Llama) and 1.97 (GPT) verbs on average, indicating that synonyms are common within the 504 target set.

- *Multi-Perspectives.* Verbs interpreted from different perspectives in a single image are identified when an image appears in multiple clusters. Both clustering results show that over 70% of images belong to more than one cluster (Row 4, Multi-Image Rate), with each image appearing in 3.61 (Llama) and 4.44 (GPT) clusters on average (Row 3, Clusters per Image), i.e., about 4 perspectives per image.

Despite differences in the number of final clusters produced by each model, the consistency observed across other metrics underscores the robustness of our evaluation framework.

Polysemy, where a verb has multiple senses, though does not affect the model evaluations, can also be analyzed through clustering results. It is identified when a verb appears in multiple clusters. As shown in Table 2, over 50% of the verbs span multiple clusters (Row 6, Multi-Verb Rate), indicating they represent different senses. On average, each verb is associated with 2 clusters (Row 5, Clusters per Verb).

4.2 Evaluation of Visual Activity Recognition Systems

Model Evaluation. The primary goal of our clustering approach is to enable more accurate evaluation of visual activity recognition systems. We validate this framework by evaluating multiple models and comparing their performance under three criteria: exact-match accuracy, WordNet-synset-based accuracy, and our cluster-based accuracy. In the WordNet-based evaluation, a prediction is considered correct if the predicted verb shares at least one synset with the gold target verb. For our cluster-based evaluation, we examine all clusters containing the target image; a prediction is deemed correct if the predicted verb appears in any of those clusters. Otherwise, it is marked incorrect. Our evaluation focuses on two types of systems:

- *Supervised Models.* We built a straightforward supervised architecture consisting of an image encoder backbone followed by a linear classification layer. We experimented with three image encoders: ResNet-50 (He et al., 2016), CLIP (Radford et al., 2021), and Llama-3.2-11B (Dubey et al., 2024). Although achieving state-of-the-art performance was not our primary objective, our Llama-11B-based model achieved strong results with a Top-1 accuracy of 56% on the imSitu test set, comparable to the current SOTA (details in Appendix C).

- *Multimodal LLMs.* To demonstrate the zero-shot capability of multimodal LLMs, we tested three models: GPT-4o mini, Llama-3.2-11B and Llama-3.2-90B. We designed two prompt settings: one providing 504 verbs to choose from (**closed**) and one without (**open**).

Table 3 presents Top-1 and Top-5 accuracy under three evaluation criteria: exact match with the gold verb, WordNet-Synset (WN-Syn) match, and our cluster-based evaluation. The top section shows that models using Llama as the backbone achieve the best performance across all metrics. While all models benefit from cluster-based evaluation, CLIP and Llama exhibit larger gains compared to ResNet, suggesting that multimodal training leads to stronger semantic understanding.

Zero-shot LLMs perform worse than supervised models under exact-match evaluation. However, their scores improve significantly with our cluster-based metric. This suggests their predictions, while not always matching the gold verb, are often semantically appropriate. We also find WordNet-based evaluation yields about a 2% gain in Top-1 accuracy compared to exact match (56.0%→58.4%). This gain reflects improvements from synonymy and aligns with the increase observed in our model accuracy improvement breakdown.

		Top-1 Accuracy			Top-5 Accuracy		
	Model	**Gold**	**WN-Syn**	**Cluster**	**Gold**	**WN-Syn**	**Cluster**
	ResNet-50	30.5	32.5	43.2	55.9	58.2	70.7
	CLIP	46.2	48.4	61.5	75.5	76.9	86.7
	Llama-3.2-11B	56.0	58.4	72.1	83.1	84.3	91.8
closed	GPT-4o mini	17.3	20.5	39.2	55.7	60.2	85.3
	Llama-3.2-11B	25.3	28.6	52.4	37.7	41.2	69.8
	Llama-3.2-90B*	24.9	27.4	-	37.0	40.2	-
open	GPT-4o mini	20.0	22.2	44.4	45.6	49.3	78.2
	Llama-3.2-11B	7.4	8.7	23.5	20.3	22.3	41.7
	Llama-3.2-90B	17.2	29.8	48.9	41.8	45.4	80.6

Table 3: Comparison of model performance using different evaluation methods. Top-1 and Top-5 accuracies are reported based on three criteria: exact match with the gold answer (Gold), WordNet synset-based evaluation (WN-Syn), and clustering-based evaluation (Cluster). The *closed* setting provides 504 verbs to the LLMs in the prompt, while the *open* setting does not. *The clusters are derived from Llama-3.2-90B (closed) and therefore not used when evaluating that same configuration.

	Model	**Gold**	**Cluster**	**Syn**	**Multi-P**
	ResNet-50	30.5	43.2	+3.9	+8.8
	CLIP	46.2	61.5	+4.9	+10.4
	Llama-11B	56.0	72.1	+5.4	+10.7
closed	GPT-4o mini	17.3	39.2	+5.0	+17.0
	Llama-11B	25.3	52.4	+6.0	+21.2
open	GPT-4o mini	20.0	44.4	+5.1	+19.4
	Llama-11B	7.4	23.5	+2.3	+13.7
	Llama-90B	17.2	48.9	+4.6	+27.1

Table 4: Top-1 accuracy improvements using cluster-based targets. *Syn*: gain from synonyms. *Multi-P*: gain from multi-perspective understanding.

Improvement Breakdown. In Table 4, we analyze the breakdown of the improvements in Top-1 accuracy in all models, examining how much of the improvement is attributed to synonyms and how much is due to different perspectives of the image. To distinguish between these cases, we identify whether the Top-1 label, which appears in the target verb clusters but is not the gold target, is in the same cluster as the gold target. If it appears in the same cluster, the prediction is classified as a synonym of the gold target. Otherwise, it is considered a description of the image from a different perspective.

Table 4 shows that models more often predict verbs from different perspectives (*Multi-P* column). For supervised classification models, addressing perspective-related challenges contributes improvements of 8% to 10%, accounting for a significant portion of the total improvement (12% to 16%). In large language models (LLMs), perspective-related adjustments yield even greater gains, with improvements as high as 27%, compared to total improvements ranging from 22% to 32%. While models also face difficulties in predicting synonyms, the impact is less pronounced than the challenges posed by differing perspectives.

Human Judgment. To assess how well different evaluation strategies reflect human judgment, we analyze alignment from two complementary perspectives: *absolute accuracy alignment* and *relative agreement in ranking*. Implementation details for the evaluation setup are provided in Appendix D.

For absolute alignment, we randomly sampled 100 images and manually annotated the correctness of Top-1 and Top-5 predictions produced by all supervised models and LLMs. Table 5 reports human-labeled accuracies alongside accuracies computed using gold labels, verb clusters, and VLM-as-a-judge (GPT-4o and GPT-4o mini). Exact-match evaluation against the gold verb systematically underestimates performance, reflecting verb semantic ambiguity, while VLM-as-a-judge consistently overestimates accuracy. In contrast, our cluster-based evaluation yields accuracy estimates that are closest to human judgments across models and settings.

Using the same 100-image subset, we further analyze relative alignment between evaluation metrics and human judgments by measuring Kendall's τ (ranking agreement) and Krippendorff's α (label agreement). Results are shown in Table 6. We compare our cluster-based evaluation against three alternatives, namely gold-label matching, VLM-as-a-judge, and a CLIPScore-based (Hessel et al., 2021)

		Top-1 Accuracy					Top-5 Accuracy				
	Model	**Gold**	**4o**	**4o-mini**	**Cluster**	**Human**	**Gold**	**4o**	**4o-mini**	**Cluster**	**Human**
	ResNet-50	34$_{(-15)}$	57$_{(+8)}$	63$_{(+14)}$	**47**$_{(-2)}$	49	58$_{(-21)}$	86$_{(+7)}$	86$_{(+7)}$	**77**$_{(-2)}$	79
	CLIP	50$_{(-20)}$	77$_{(+7)}$	84$_{(+14)}$	**64**$_{(-6)}$	70	79$_{(-15)}$	**94**$_{(0)}$	98$_{(+4)}$	89$_{(-5)}$	94
	Llama-3.2-11B	62$_{(-20)}$	90$_{(+8)}$	88$_{(+6)}$	**79**$_{(-3)}$	82	81$_{(-15)}$	100$_{(+4)}$	99$_{(+3)}$	**94**$_{(-2)}$	96
closed	GPT-4o mini	16$_{(-40)}$	75$_{(+19)}$	78$_{(+22)}$	**39**$_{(-17)}$	56	52$_{(-35)}$	98$_{(+11)}$	98$_{(+11)}$	**85**$_{(-2)}$	87
	Llama-3.2-11B	16$_{(-31)}$	**52**$_{(+5)}$	**52**$_{(+5)}$	39$_{(-8)}$	47	20$_{(-38)}$	83$_{(+25)}$	82$_{(+24)}$	**61**$_{(+3)}$	58
open	GPT-4o mini	18$_{(-43)}$	80$_{(+19)}$	77$_{(+16)}$	**48**$_{(-13)}$	61	44$_{(-38)}$	95$_{(+13)}$	97$_{(+15)}$	**76**$_{(-6)}$	82
	Llama-3.2-11B	6$_{(-20)}$	38$_{(+12)}$	36$_{(+10)}$	**23**$_{(-3)}$	26	16$_{(-24)}$	74$_{(+34)}$	84$_{(+44)}$	**41**$_{(+1)}$	40
	Llama-3.2-90B	12$_{(-43)}$	75$_{(+20)}$	70$_{(+15)}$	**55**$_{(0)}$	55	34$_{(-47)}$	95$_{(+14)}$	94$_{(+13)}$	**80**$_{(-1)}$	81

Table 5: Comparison of Top-1 and Top-5 accuracy on 100 randomly sampled test images. Accuracy is measured using gold labels (exact match), GPT-4o and GPT-4o mini as judges, verb clusters, and human annotations. Subscripts show differences from human labels, and **bold** marks the closest value.

Metric	**Kendall's τ**		**Krippendorff's α**	
	Top-1	**Top-5**	**Top-1**	**Top-5**
Gold (exact match)	69.1	78.6	39.5	37.9
CLIPScore-based	57.1	78.6	33.4	47.6
Cluster (ours)	76.4	85.7	56.7	**60.7**
GPT-4o as judge				
r1	90.9	76.4	**63.2**	44.6
r2	76.4	92.9	62.4	45.8
r3	**98.2**	88.9	47.2	30.5
r4	**98.2**	**92.9**	58.4	54.0
GPT-4o mini as judge				
r1	92.9	90.9	56.7	35.0
r2	90.9	90.9	58.7	33.6
r3	85.7	81.5	44.1	20.8
r4	92.9	61.8	56.4	28.4

Table 6: Agreement between evaluation metrics and human judgments on the 100-image subset. We report Kendall's τ and Krippendorff's α, with all values scaled by 100 for readability. For GPT-based judges, rounds r1–r2 use the same prompt as in Table 5, while r3–r4 use varied prompts to assess prompt sensitivity.

method. For the CLIPScore criterion, we compute the CLIPScore between the predicted verb and the image, and deem the prediction correct if this score is greater than or equal to the CLIPScore between the gold verb and the same image. For VLM-as-a-judge, we report four independent rounds, each a separate run of the judge on the same images. Rounds 1-2 use the same prompt as in Table 5 to assess sampling stochasticity. Rounds 3-4 use alternative prompts to evaluate robustness to prompt variation.

Our cluster-based evaluation achieves substantially higher agreement with human judgments than both exact match and the CLIPScore-based method across Top-1 and Top-5 evaluations. In contrast, VLM-as-a-judge exhibits notable instability, with agreement varying across repeated runs using the same prompt and fluctuating further when prompts are changed. Although GPT-4o and GPT-4o mini often preserve relative model ordering and thus attain high Kendall's τ despite overestimating absolute performance, Krippendorff's α offers a clearer view of absolute alignment by showing that our cluster-based evaluation aligns more reliably with human labels and exhibits lower variability.

Robustness to Cluster Count. We evaluate the robustness of our cluster-based framework by systematically varying cluster counts while holding other parameters constant. By re-assessing visual activity recognition models under these different cluster configurations, we examine how evaluation outcomes depend on cluster granularity. As shown in Figure 4, models' Top-1 accuracy demonstrates stable evaluation patterns across cluster counts. We maintain methodological consistency with our primary evaluation (Section 4.2), using K-Means clustering of Llama-3.2-90B's closed-prompt verb predictions as our basis.

As expected, our analysis reveals a consistent inverse correlation between cluster granularity and model accuracy. This expected pattern emerges because finer clusters impose stricter semantic criteria, progressively limiting the number of valid target verbs per image. The accuracy decline stabilizes at approximately 655 clusters, which serves as the configuration for our main evaluations. To properly benchmark these results, we implement a baseline (visually denoted by a dashed horizontal line) using unprocessed Llama-3.2-90B outputs (closed) as direct reference data - the same outputs used for cluster generation. We count predictions as correct when they match any of the reference verbs. The performance gap between the baseline and our cluster-based evaluation demonstrates our

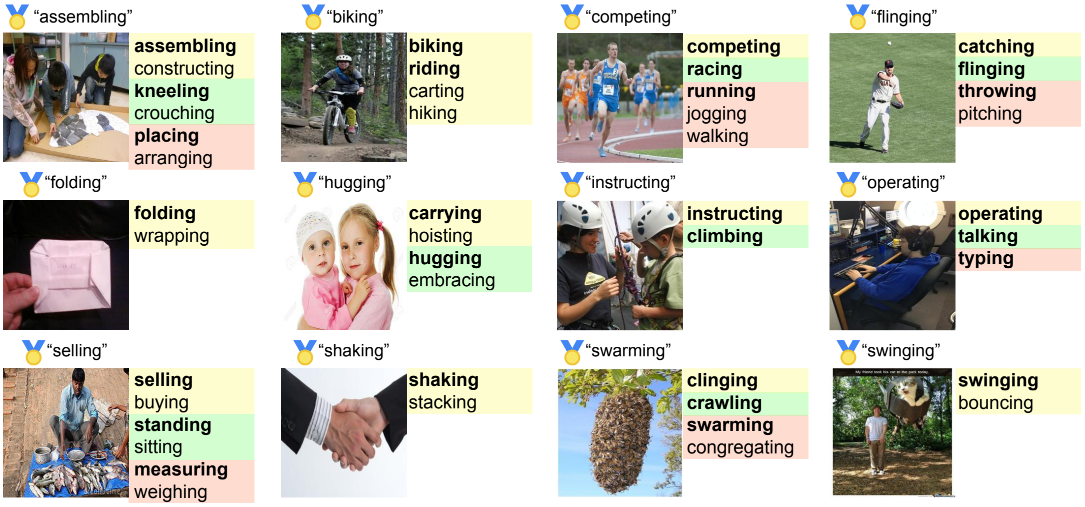


Figure 3: Qualitative examples of our clustering output. Each mini-panel shows: the imSitu gold verb (medal icon), Llama-3.2-90B predictions in **bold**, and additional verbs grouped in the same sense cluster. Different background colors indicate distinct clusters associated with the image.

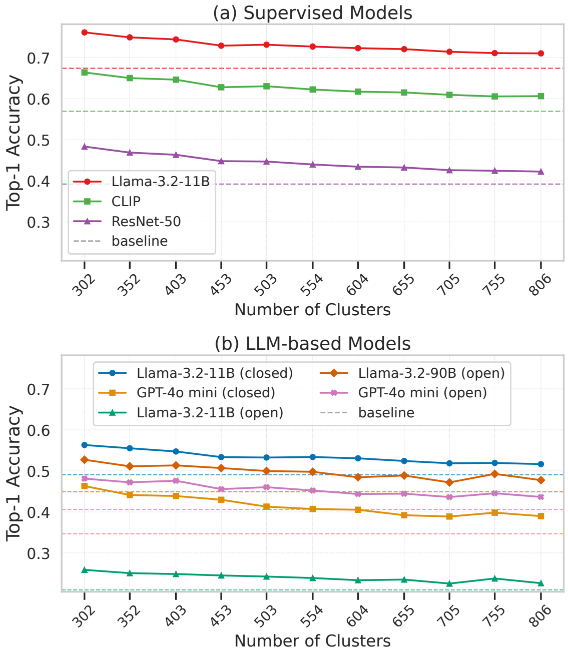


Figure 4: Top-1 accuracy of different models based on our clustering results. The dashed baselines represent the accuracy of the models when we use only the Llama-3.2-90B (closed) outputs as target verbs (no clustering), i.e., model's prediction verb is considered correct when it matches any of the targets.

method's crucial advantage: its inherent capacity to recognize linguistically valid variations, including synonymous expressions and semantically related verb forms, while maintaining rigorous evaluation standards.

Summary. We evaluate both supervised models and LLMs using our cluster-based framework, showing improved accuracy driven by both synonym resolution and multi-perspective understanding. The evaluation aligns more closely with human judgment, and robustness analysis confirms consistent accuracy trends across different clustering settings.

4.3 Example of our Clustering Results

To qualitatively demonstrate the results of our clustering, we present Figure 3. The 12 examples shown are randomly selected from the test set, each with its gold label (illustrated by the medal emoji), inference predictions (in **bold**), and other verbs within the same cluster. Different clusters associated with the same image are distinguished by background colors.

Using the first image as an example, Llama-3.2-90B predicted *assembling*, *kneeling*, and *placing*. After our clustering steps, *assembling* was grouped with *constructing*, *kneeling* with *crouching*, and *placing* with *arranging*. The fact that this image is associated with three distinct clusters supports the interpretation that multiple activities are occurring simultaneously: *assembling* reflects the overall task, *kneeling* describes the physical posture, and *placing* captures the fine-grained action of positioning components. While the three inference verbs in this case correspond to different activities, there are examples where multiple predictions express the same underlying action. For instance, in the *biking* image shown in the second column, Llama predicted both *biking* and *riding*, which were correctly grouped into a single cluster by our method.

	Statistics	Img-Verb$_{gold}$	Img-Cap$_{gold}$	GPT Adversarial		Random Choice	
				Img-Cap$_{adv}$	Cap$_{gold}$-Cap$_{adv}$	Img-Cap$_{rand}$	Cap$_{gold}$-Cap$_{rand}$
Dev	Mean	23.92	27.41	26.03	92.59	26.20	94.03
	Std	2.99	3.53	3.76	3.59	3.81	4.38
Test	Mean	23.93	27.41	26.04	92.57	26.18	93.95
	Std	2.98	3.50	3.73	3.92	3.79	4.43

Table 7: CLIPScores (scaled by 100) for image-caption alignment using gold captions (Img-Cap$_{gold}$), GPT adversarial captions (Img-Cap$_{adv}$), and random-verb captions (Img-Cap$_{rand}$). We also report caption-caption CLIPScores (Cap$_{gold}$-Cap$_{adv}$, Cap$_{gold}$-Cap$_{rand}$) to quantify textual similarity under these perturbations, and include Img-Verb$_{gold}$, which measures the CLIPScore between each image and its gold verb as a baseline reference.

While our method performs well in most cases, minor errors could still arise in the pipeline, both from the Llama model's inference and from our clustering. For example, in the *biking* image, *carting* was incorrectly included in the same cluster, even though it does not accurately describe the scene. In the *selling* image at the end of the first column, the Llama model labeled the activity as *standing*, while the person is clearly *sitting*. Interestingly, our clustering partially corrects this by including *sitting* in the target cluster. A more detailed error analysis is provided in Appendix E.

4.4 Analysis of CLIPScore Performance

While discrete gold labels often miss the nuances of visual activity recognition, an open question is whether score-based image-text alignment metrics can provide a better alternative. We study this question using CLIPScore (Hessel et al., 2021), which measures semantic similarity via CLIP embeddings.

A natural instantiation is verb-only alignment, where we compare an image directly to a *single verb* by computing the CLIPScore between the image and the predicted verb. To assess this setting, we compute CLIPScores between each image and all 504 imSitu verbs, rank verbs by score, and evaluate Top-1 and Top-5 accuracy against the gold labels. CLIPScore reaches a Top-1 accuracy of $\sim$25%, which is plausible given the multi-perspective nature of images, but Top-5 remains below 50%, suggesting difficulty in reliably retrieving even human-validated labels. We conjecture that this limitation stems in part from CLIP being trained primarily on caption-like text rather than isolated action words, so the verb signal is weak and easily confounded.

Motivated by this, we next test whether CLIPScore behaves more reliably in a caption setting. Our experiment on imSitu (Table 7) examines how verb changes within otherwise identical captions affect CLIPScore. Using the labeled verb and its semantic roles, we generate correct captions for each image (see Appendix F for details). From each caption, we construct two modified variants by replacing the verb with an implausible alternative using GPT-4o (e.g., *The man kneads clay at the garage.* $\rightarrow$ *The man eats clay at the garage.*) or by substituting it with a random verb from the imSitu verb list. While correct captions yield slightly higher CLIPScores, the differences are minimal even for semantically incorrect substitutions. The corresponding caption-caption CLIPScores remain high ($>$ 92%), and Img-Cap$_{gold}$ is only slightly higher than Img-Verb$_{gold}$, suggesting that adding caption context only weakly strengthens the verb signal. Overall, this points to a limitation of CLIPScore for action-focused evaluation, because contextual words dominate the similarity signal and changes to the verb have little effect, a tendency also reported in textual-only settings (Garcia, 2021).

5 Conclusion

In this work, we address the challenge of evaluating visual activity recognition systems by developing a two-step clustering framework that accounts for inherent ambiguities in verb semantics and image interpretation. Our analysis reveals that images in the imSitu dataset frequently involve both synonymous verbs describing the same activity and multiple valid interpretations from different perspectives. Through evaluating multiple recognition systems, including both supervised models and zero-shot multimodal LLMs, we show that our cluster-based evaluation yields higher accuracy scores compared to exact-match evaluation, which better reflects these models' capabilities in understanding visual activities, as validated by our manual analysis.

Limitations

While our results demonstrate the effectiveness of cluster-based evaluation for visual activity recognition, several limitations remain. First, our experiments are conducted solely on the imSitu dataset. Although imSitu provides a well-established benchmark with rich verb annotations, evaluating the framework on additional datasets would further strengthen its generality. Importantly, our approach is not tied to imSitu-specific annotations. To adapt the framework to other datasets, such as VidSitu or datasets with verb-phrase labels, one can (i) collect model-generated verb or verb-phrase predictions for each visual instance, (ii) embed image-text pairs using a joint vision-language model, (iii) cluster the resulting representations using the same data-driven criteria, and (iv) evaluate model predictions against the resulting clusters. For verb-phrase labels, phrases can be embedded directly or decomposed into head verbs with modifiers, enabling the same clustering procedure to be applied. While this process requires dataset-specific preprocessing and validation, it provides a principled starting point for constructing visually grounded evaluation criteria beyond exact-match labels.

Second, our clustering relies on verb outputs generated by large multimodal models (e.g., Llama-3.2-90B and GPT-4o mini), rather than exhaustive human annotation. Although we manually inspected random samples to validate output quality, more comprehensive human evaluations or precision–recall analyses would further strengthen this component. In addition, our current framework treats all verbs within a cluster as equally valid, without modeling finer-grained contextual preferences. Incorporating weighted or context-dependent relationships between verbs is an important direction for future work.

Finally, our method introduces a nontrivial one-time computational cost for candidate generation, embedding, and clustering, and can be more expensive than a minimal VLM-as-a-judge baseline when evaluating a single system once. However, this cost is largely offline and amortized, since the procedure yields a fixed set of cluster-based targets per image that can be reused across experiments without additional VLM calls. In contrast, VLM-as-a-judge incurs repeated inference for each new system output, so its cost scales with the number of evaluated models, prompts, or checkpoints.

Acknowledgements

We thank the CincyNLP group for helpful discussions, and the anonymous reviewers for their valuable feedback and constructive suggestions.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015)*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*.

Junhyeong Cho, Youngseok Yoon, and Suha Kwak. 2022. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*.

Thilini Cooray, Ngai-Man Cheung, and Wei Lu. 2020. Attention-based context aware reasoning for situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL 2021)*.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. AVA: A video dataset of spatio-temporally localized atomic visual actions. In

2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018).

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)*.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL 2021 (Findings of ACL-IJCNLP 2021)*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Tianyu Jiang and Ellen Riloff. 2023. Exploiting commonsense knowledge about objects for visual activity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023 (Findings of ACL 2023)*.

Dieu Thu Le, Raffaella Bernardi, and Jasper Uijlings. 2013. Exploiting language models to recognize unseen actions. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval (ICMR 2013)*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. UMIC: An unreferenced metric for image captioning via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL 2021)*.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL 2002)*.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference (ECCV 2020)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML 2021)*.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. Semeval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Matteo Ruggero Ronchi and Pietro Perona. 2015. Describing common human visual actions in images. In *Proceedings of the British Machine Vision Conference (BMVC 2015)*.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Debaditya Roy, Dhruv Verma, and Basura Fernando. 2024. Clipsitu: Effectively leveraging clip for conditional predictions in situation recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2024)*.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*.

Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. 2024. HARDVS: Revisiting human activity recognition with dynamic vision sensors. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2024)*.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. 2011. Human action recognition by learning bases of action attributes and parts. In *2011 International Conference on Computer Vision (ICCV 2011)*.

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)*.

A Prompting LLMs

As we aim to evaluate the multiple perspectives of the images, one of the key challenges is obtaining comprehensive descriptions that include all perspective verbs associated with an image. To acquire this information, we employ state-of-the-art (SOTA) multimodal large language models (LLMs) to label the images. We utilize Llama-3.2-11B and Llama-3.2-90B (Dubey et al., 2024), both widely regarded as among the best open-source models available. Additionally, we query GPT-4o mini (Hurst et al., 2024) through the OpenAI API. While GPT-4o is considered superior, labeling over 100k images with it is beyond our budget constraints.

For this task, we experimented with two prompting strategies. The first is an open prompt, which asks the model to output the top 5 verbs that describe the image. The second is a closed prompt, where the 504 target verbs are provided, and the model is asked to select all verbs that describe the activity in the image. Both prompts are detailed below:

Open Prompt:

```
What are some verbs that describe what
is happening in this image?
Answer only with comma separated verbs
in the gerund form (they end in 'ing').
Do not include more than 5.
```

Closed Prompt:

```
You are provided with a list of 504
verbs: eating, walking, piloting, ...
Identify and list all verbs from this
set that accurately describe the
activity depicted in the image.
Respond with the verbs only, separated
by commas.
```

Since we focus on the 504 verbs in the imSitu dataset, we filter the output of the open prompt by matching it to this list. However, we observed that while the LLMs often generated reasonable replies with the open prompt, the verb distribution differed significantly. After filtering, as much as 50% of the images received no valid replies within the 504 list for Llama-3.2-11B. Consequently, we use the closed prompt results as the primary gold labels for the clustering steps to identify ambiguities.

Llama-3.2-11B often produced uninformative replies, such as “I cannot assist with this task,” or repeated all 504 verbs in its response. Consequently, we rely exclusively on the results from the closed prompt of Llama-3.2-90B and GPT-4o mini for the clustering steps, omitting Llama-3.2-11B due to its lower reliability.

B Clustering Evaluation Metrics

To evaluate the quality of the clustering results, we employ three widely-used metrics: Silhouette score, Calinski-Harabasz index, and Purity score. Each metric provides a complementary view of clustering performance by quantifying different aspects of cluster quality, including cohesion, separation, and alignment with ground-truth labels. Both the Silhouette score and Calinski-Harabasz index are unsupervised metrics that measure the balance between within-cluster compactness and between-cluster separation. The key difference lies in their focus: the Silhouette score evaluates pairwise distances between points, while the Calinski-Harabasz index assesses deviations of points from their cluster centroids.

Silhouette score: The Silhouette score quantifies how well-separated and cohesive the clusters are. For a data point i, the Silhouette score is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where $a(i)$ is the average distance from i to all other points within the same cluster, and $b(i)$ is the smallest average distance from i to points in any other cluster. In our evaluation, we compute the Silhouette score using cosine distance to better capture semantic similarities in the high-dimensional embedding space. The overall Silhouette score is the mean of $s(i)$ over all data points. Higher values, closer to 1, indicate well-separated and compact clusters.

Calinski-Harabasz Index: The Calinski-Harabasz index, also called the Variance Ratio Criterion, measures how well-defined the clusters are by comparing how far apart the clusters are to how tightly packed the points are within each cluster. It is calculated as:

$$CH = \frac{\text{trace}(B_k)}{\text{trace}(W_k)} \cdot \frac{N - k}{k - 1},$$

where B_k and W_k are scatter matrices that quantify the dispersion of data points. Specifically, B_k (between-cluster scatter matrix) measures how spread out the cluster centroids are from the overall data centroid, capturing the separation between clusters. In contrast, W_k (within-cluster scatter matrix) measures how tightly data points are grouped around their respective cluster centroids, reflecting the compactness of individual clusters. Additionally, N represents the total number of data points, and k denotes the number of clusters.

To make sure the distance calculations align with cosine similarity, we normalize all embeddings to have a length of 1 before clustering. A higher Calinski-Harabasz index means the clusters are well-separated from each other and the points within each cluster are tightly grouped.

Purity Score: Since there is no gold standard to evaluate our clustering results, we compute the Purity score using WordNet Synsets (Miller, 1995). For each cluster, we identify all synsets that contain at least one verb from the cluster and calculate the Purity score based on the maximum overlap between the cluster and any synset. To ensure meaningful evaluation, we calculate Purity only for clusters containing at least two verbs, as mapping single-verb clusters to the correct synset for comparison is ambiguous. The Purity score is defined as:

$$\text{PuS} = \frac{\sum_{|\text{Cluster}i|>1} \max j |\text{Cluster}_i \cap \text{Synset}j|}{\sum_{|\text{Cluster}_i|>1} |\text{Cluster}_i|},$$

where $|\text{Cluster}_i|$ is the size of cluster i, and $|\text{Cluster}_i \cap \text{Synset}_j|$ represents the overlap between cluster i and synset j. This metric approximates the semantic coherence of clusters by measuring their alignment with WordNet synsets.

C Supervised Classification Model

To evaluate the performance of the classification model using our clusters, we train three different supervised classification models. We leverage three image encoders: CLIP (Radford et al., 2021), ResNet-50 (He et al., 2016), and Multimodal Llama-11B (Dubey et al., 2024)—for verb prediction on the imSitu training set. Each backbone processes the input and generates embeddings of varying dimensions:

- **ResNet-Based Model:** Uses ResNet-50, trained on ImageNet-1k (Version 1), to generate 2048-dimensional visual embeddings.

- **CLIP-Based Model:** Utilizes the Vision Transformer architecture, ViT-B/32 (Base, Patch Size 32), to produce 512-dimensional image embeddings.

- **Llama-11B-Based Model:** The model processes the input image through its vision encoder, while the text stream contains only a special `<|image|>` token. Both the image and text streams are passed through the language model using the cross-attention mechanism. The final token from the last layer serves as a 4096-dimensional embedding.

Model	Top1-G	Top1-C	Top5-G	Top5-C
CRF (Yatskar et al., 2016)	0.32	-	0.59	-
CLIP-Event (Li et al., 2022)	0.46	-	-	-
ARF (Jiang and Riloff, 2023)	0.47	-	-	-
ClipSitu (Roy et al., 2024)	0.58	-	0.86	-
ResNet	0.31	0.43	0.56	0.71
CLIP	0.46	0.62	0.75	0.87
Llama	0.56	0.72	0.83	0.92

Table 8: Performance comparison between our supervised models and previous state-of-the-art on imSitu test set.

In all three models, the backbone parameters are kept frozen, and a trainable linear classification head is added on top. This classification head projects the embeddings to an output dimension of 504, representing the number of verbs in the dataset. The outputs are processed through a softmax layer and trained using cross-entropy loss, with Stochastic Gradient Descent (SGD) as the optimizer.

Early works (Yatskar et al., 2016; Pratt et al., 2020) relied on CNN-based backbones for image encoding, establishing a foundation for visual activity recognition. Recent advancements have introduced CLIP-based models (Li et al., 2022; Roy et al., 2024) and transformer-based architectures (Jiang and Riloff, 2023), leading to significant improvements in performance. Our ResNet-based and CLIP-based models closely match the performance reported in these prior studies, as shown in Table 8, confirming the robustness and effectiveness of our implementations.

While achieving state-of-the-art (SOTA) performance was not the primary objective of this study, our Llama-11B-based model demonstrates strong results, achieving a Top-1 accuracy of 56% on the imSitu test set. This performance approaches the current SOTA (Roy et al., 2024) for verb prediction, underscoring the potential of large-scale language models in this domain.

D LLM-as-Judge

Recently, LLM-as-Judge has emerged as a popular paradigm for automatic evaluation, offering scalability but raising concerns about reliability and alignment with human judgment. To assess the viability of this approach, we conducted two complementary experiments on the same 100-image subset used in our human judgment analysis, using GPT-4o and GPT-4o mini as judges. First, we evaluate model performance via an accuracy-based comparison by querying whether model predictions correctly describe the image. Second, we analyze agreement with human judgments using Kendall's τ and Krippendorff's α, enabling a direct comparison of ranking consistency and label-level alignment.

D.1 Accuracy-Based Evaluation

For the accuracy-based evaluation, we query the LLM judges using the following binary prompts for Top-1 and Top-5 predictions.

Top 1:

```
Does verb 'VERB' describe the image?
Respond with 'yes' or 'no'.
```

Top 5:

```
Do any verbs in 'VERB1, VERB2, ...'
describe the image?
Respond with 'yes' or 'no'.
```

We found that both models tended to overestimate alignment with human-labeled correctness, and their judgments varied substantially across model outputs. The discrepancy between GPT-4o and GPT-4o mini further highlights the inherent instability of LLM-as-Judge approaches in this setting.

D.2 Agreement-Based Evaluation

Beyond absolute accuracy, we evaluate alignment between LLM-based judges and human judgments using two complementary agreement measures. Kendall's τ measures agreement in relative ranking between two orderings over the same set of models. Intuitively, it asks whether pairs of models are ranked in the same order by two evaluation methods. For any pair of models, the pair is considered *concordant* if both rankings agree on which model performs better, and *discordant* if they disagree. Kendall's τ is then defined as

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)},$$

where N_c and N_d denote the numbers of concordant and discordant model pairs among n models. A higher value of τ indicates stronger consistency in relative model ordering.

Krippendorff's α measures *label-level agreement* between LLM judgments and human annotations across individual instances. It is defined as

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o denotes the observed disagreement between annotators and D_e denotes the expected disagreement under chance. Higher values of α indicate closer alignment in absolute labeling decisions.

Using the same 100-image subset, we compute both metrics for multiple runs of GPT-4o and GPT-4o mini. Specifically, rounds r1 and r2 use the same prompt as defined in Section D.1, allowing us to assess variability across repeated runs. Rounds r3 and r4 introduce prompt variations to evaluate sensitivity to prompt formulation, and the corresponding prompts are shown below.

Round 3 Prompts:
Top-1:

```
Is the activity 'TOP_1' happening
in the image?
Respond with 'yes' or 'no'.
```

Top-5:

```
Is any activity in 'VERB1, VERB2, ...'
happening in the image?
Respond with 'yes' or 'no'.
```

Round 4 Prompts:
Top-1:

```
Determine whether the verb 'TOP_1' is
visually supported by the image.
Respond with 'yes' or 'no'.
```

Top-5:

```
Determine whether any verb in
'VERB1, VERB2, ...'
is visually supported by the image.
Respond with 'yes' or 'no'.
```

This design enables a direct analysis of stability in LLM-as-Judge evaluations and facilitates comparison with our cluster-based evaluation in terms of both ranking consistency and label-level reliability.

D.3 Discussion

Our results show that off-the-shelf LLM-as-Judge both overestimates model accuracy relative to human judgments and exhibits substantial instability under prompt perturbations, with agreement varying across repeated runs and prompt formulations. While more advanced prompting techniques, such as Chain-of-Thought or in-context examples, may improve reliability, our clustering approach provides a more stable and interpretable alternative by producing explicit verb sense groupings that support both dataset-level and instance-level analysis. In contrast, LLM-as-Judge methods yield binary or probabilistic judgments without an explicit semantic structure, a limitation that has been increasingly noted in recent studies (Chen et al., 2024; Thakur et al., 2024).

E Error Analysis

While our clusters have proven effective in capturing verb semantics, marginal errors still occur throughout the pipeline. We conduct an error analysis by categorizing these errors into two main types: *inference error* and *clustering error*, with illustrative examples shown in Figure 5. Each example includes the gold label, inference predictions (in **bold**), and other verbs within the same cluster. Different clusters corresponding to the same image are distinguished by background color.

Inference error arises during the generation of `<image,verb>` pairs and typically occurs when the predicted activity does not accurately or completely reflect the image content. As shown in the first column of Figure 5, the verb *typing* is an incorrect prediction for the image labeled *shouting*, and while *photographing* may be loosely related to the image labeled *spying*, the inclusion of *peeing* is clearly irrelevant. Another form of inference error occurs when plausible activities are omitted. As shown in the second column, *serving* would be an equally valid label for the *pouring* image, and the man in the *talking* image is clearly also *phoning*.

Clustering error is also two-fold. The first type occurs when semantically unrelated or loosely related activities are grouped together. As shown in the third column of Figure 5, while *skiing* and *skating* are clearly related, *swimming*, though also a sport, is less relevant. Similarly, *sitting* and *standing* often co-occur but are semantically distinct: for instance, *sitting* can describe the *driving* image, while *standing* cannot. The second type of clustering error involves cases where multiple related clusters could be merged. Although this type is less critical for evaluation, we show examples in the last column. In the *shopping* image, *shopping* and {*buying*, *selling*} are closely related and frequently co-occur, suggesting they could belong to the same cluster. Similarly, *dining* and *eating* could be grouped together to better describe the *dining* image.

F Analysis on CLIPScore

This section presents further details regarding the CLIPScore analysis.

With context: To perform the evaluation, we generate both correct and modified captions for each image. In the imSitu dataset, each verb is associated with an abstract template and a predefined set of semantic roles. For example, the verb *soaring* is linked to the abstract template *an AGENT soars in a PLACE*, where *AGENT* and *PLACE* denote semantic roles. These roles are annotated for each image labeled with the corresponding verb. To obtain the correct caption, we prompt GPT-4o mini to generate a natural language sentence. An example prompt is shown below:

```
Given the abstract structure 'an AGENT
soars in a PLACE' and the following
components:
'place': 'sky',
'agent': 'bird',
generate a natural sentence.
Reply with only the sentence.
```

After obtaining the correct caption, we generate modified captions using two approaches. The first involves replacing the verb in the correct caption with a randomly selected verb from the imSitu verb list. The second approach prompts GPT-4o to substitute the verb with an implausible alternative. The prompt used for this transformation is as follows:

```
You are given the following sentence:
'{correct caption}'
```

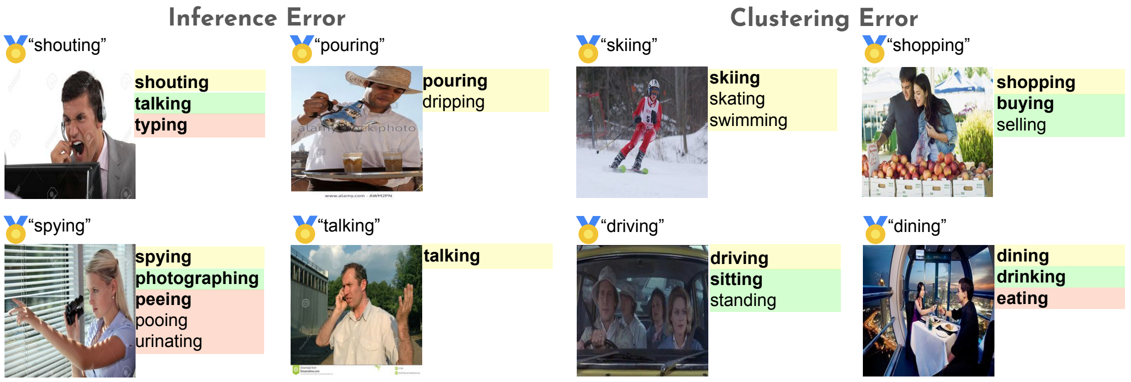


Figure 5: Error analysis of our clustering pipeline. Inference predictions are shown in **bold**, and different clusters for the same image are distinguished by background colors.

```
Can you change the verb in the sentence
to an unreasonable one without
changing other parts of the sentence?
Reply with only the new sentence.
```

Once both the correct and modified captions are generated, we compute the cosine similarities between their CLIP embeddings. The results are presented in Table 7.

	Top 1	**Top 5**
Train	0.25	0.48
Test	0.26	0.49
Dev	0.26	0.49

Table 9: Accuracy on the imSitu dataset using CLIPScore for verb ranking.

Without context: To further study the ability of CLIPScore for evaluating image-activity alignment in the absence of context, we compute CLIP similarity scores between each image and all 504 verbs in the imSitu dataset. The verbs are then ranked by their CLIPScore, and we assess Top-1 and Top-5 prediction accuracy against the gold-standard labels. The results, presented in Table 9, show that CLIPScore achieves a Top-1 accuracy of approximately 25% (i.e., 25% of the images receive their highest similarity score with the gold-standard verb), while the Top-5 accuracy remains below 50%. These results are consistent across the train, dev, and test splits.

G Clustering Algorithm

Algorithm 1 Two-Step Clustering Framework with Silhouette Optimization

```
Input: A dataset of <image,verb> pairs $\mathcal{D}$, a list of 504 target verbs $\mathcal{V}$, a clustering algorithm $\mathcal{A}$, a candidate list of cluster numbers $\mathcal{K}$ for Step 1, and a candidate list of ratios $\mathcal{R}$ for Step 2.
Output: A set of final clusters $\mathcal{C}_{\text{final}}$.
Step 1: Same-Verb Clustering
for each verb $v \in \mathcal{V}$ do
    Collect all images $\mathcal{I}_v$ associated with $v$ from $\mathcal{D}$.
    Generate embeddings $\mathcal{E}_v$ for $\mathcal{I}_v$ using a multimodal language model.
    Initialize best_score $\leftarrow -\infty$ and best_clusters $\leftarrow \emptyset$.
    for each $k \in \mathcal{K}$ do
        Apply $\mathcal{A}$ to $\mathcal{E}_v$ with $k$ clusters to obtain clustering result $\mathcal{C}_v^{(k)}$.
        Compute the Silhouette score silhouette($\mathcal{C}_v^{(k)}$).
        if silhouette($\mathcal{C}_v^{(k)}$) $>$ best_score then
            Update best_score $\leftarrow$ silhouette($\mathcal{C}_v^{(k)}$).
            Update best_clusters $\leftarrow \mathcal{C}_v^{(k)}$.
        end if
    end for
    Save best_clusters as the optimal clustering for $v$.
end for
Store all optimal clusters from Step 1 as intermediate clusters $\mathcal{C}_{\text{step1}}$.
Step 2: Cross-Verb Clustering
Compute the total number of verbs as $|\mathcal{V}| = 504$.
Initialize best_score $\leftarrow -\infty$ and final_clusters $\leftarrow \emptyset$.
for each ratio $r \in \mathcal{R}$ do
    Compute candidate cluster number $k_r = \text{int}(504 \times r)$.
    Apply $\mathcal{A}$ to the average embeddings of $\mathcal{C}_{\text{step1}}$ with $k_r$ clusters to obtain clustering result $\mathcal{C}_{\text{final}}^{(k_r)}$.
    Compute the Silhouette score silhouette($\mathcal{C}_{\text{final}}^{(k_r)}$).
    if silhouette($\mathcal{C}_{\text{final}}^{(k_r)}$) $>$ best_score then
        Update best_score $\leftarrow$ silhouette($\mathcal{C}_{\text{final}}^{(k_r)}$).
        Update final_clusters $\leftarrow \mathcal{C}_{\text{final}}^{(k_r)}$.
    end if
end for
Return: final_clusters.
```