VDAct 2.0: Scaling Video-Grounded Dialogue for Event-driven Activity Understanding with LLM-Assisted Filtering

Wiradee Imrattanatrai[1], **Masaki Asada**[1], **Kimihiro Hasegawa**[2], **Ken Fukuda**[1], **Teruko Mitamura**[2]
[1]National Institute of Advanced Industrial Science and Technology (AIST)
[2]Language Technologies Institute, Carnegie Mellon University
{wiradee.imrattanatrai, masaki.asada, ken.fukuda}@aist.go.jp

Abstract

We present VDAct 2.0, an enhanced benchmark for video-grounded dialogue that builds upon the original VDAct by expanding dialogue coverage and introducing a scalable LLM-assisted filtering pipeline to ensure high-quality, grounded QA pairs. VDAct 2.0 comprises 6,356 human-annotated dialogues with a total of 63,958 turns, grounded in 2,975 household activity videos, with undesirable dialogue turns systematically identified and removed. To achieve this, we design a trigger-based quality framework and calibrate a panel of high-agreement LLMs through human-in-the-loop calibration, allowing scalable QA-turn-level filtering. We benchmark a wide range of pretrained and fine-tuned models, both open-source and proprietary, across standard text generation metrics and LLM-based evaluations. The results highlight both recent advances and remaining challenges in video-grounded dialogue modeling, positioning VDAct 2.0 as a high-fidelity testbed for evaluating and advancing multimodal reasoning in interactive settings.

Keywords: Video-ground dialogue, Multimodal systems, LLM-assisted filtering.

1. Introduction

The task of video-grounded dialogue, which requires an agent to generate responses based on video content, has emerged as an important direction in multimodal artificial intelligence (Alamri et al., 2019; Pasunuru and Bansal, 2018; Wang et al., 2023b). This task moves beyond static image understanding to engage with dynamic, temporal events, demanding advanced system capabilities in multimodal comprehension, temporal reasoning, and contextual interpretation (Imrattanatrai et al., 2025). Successfully addressing this challenge is crucial for developing systems that can understand and discuss complex activities, thereby expanding their capacity for nuanced and accurate interaction in real-world scenarios. Therefore, the creation of new datasets for video-grounded dialogue provides a valuable foundation for developing and evaluating more advanced models.

However, developing robust models for this task presents a fundamental challenge that reflects a broader evolution in AI research. As the field shifts from a model-centric to a data-centric paradigm (Mazumder et al., 2023; Zha et al., 2025), where data quality is increasingly recognized as the primary bottleneck to performance, video-grounded dialogue resources face two interrelated limitations. First is the issue of scale. In contrast to their image-based counterparts, which comprise millions of dialogue turns (Das et al., 2017; Feng et al., 2023), video-grounded dialogue datasets remain limited (Alamri et al., 2019; Wang et al., 2023b), especially those involving long-range videos. Such settings are vital for realistic multimodal understanding, yet their data scarcity (Imrattanatrai et al., 2025) highlights the need for further scaling of long-range, event-driven dialogue resources. A more fundamental challenge lies in ensuring data quality. While human-generated dialogue data is crucial for capturing natural language variability, scaling its volume can introduce non-grounded, speculative, or stylistically inconsistent content. Such noise can hinder effective model learning (Cai et al., 2020; Gupta et al., 2022), as well as increase uncertainty and ambiguity in model generation. Analyses of existing dialogue datasets, such as VDAct (Imrattanatrai et al., 2025), reveal that a non-trivial portion of QA pairs contain patterns such as speculative phrasing (*e.g.,* "I think he might..."), inferences about mental states (*e.g.,* "he seems happy"), or other non-factual expressions. While natural in conversation, such expressions can introduce noise during model training.

To address these challenges, we propose a scalable framework for constructing high-quality video-grounded dialogue datasets. Rather than manual cleaning, we introduce an automated filtering process that leverages high-agreement Large Language Models (LLMs). This design mitigates biases inherent in any single model and ensures judgment reliability through a human-in-the-loop calibration. The framework systematically removes speculative or ungrounded utterances while preserving factually-grounded dialogue, thereby achieving data scaling and quality enhancement simultaneously. Its design is broadly applicable to other multimodal dialogue datasets with minimal adaptation.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2653–2666
11–16 May 2026.

Building on the existing dataset, VDAct, we incorporated additional human annotations and then applied our automated filtering pipeline across the full data collection. The result is VDAct 2.0 [1]: a new resource that is both larger than the original dataset and of verifiably higher quality. This new dataset enables more reliable training and more accurate evaluation of Vision Language Models (VLMs).

This paper offers the following contributions: (1) we introduce a framework for enhancing dialogue dataset quality by filtering undesirable dialogue patterns using a calibrated panel of high-agreement LLMs, (2) we release VDAct 2.0, a larger-scale, higher-quality resource, by first expanding dialogues, then applying an LLM-based filtering, and (3) we provide initial benchmarks of state-of-the-art VLMs on VDAct 2.0 and compare them with the model performance on the existing VDAct.

2. Related Works

Video Dialogue Datasets Early multimodal dialogue research began with image-grounded settings, exemplified by the Visual Dialog dataset (Das et al., 2017). Recent large-scale efforts such as MMDialog (Feng et al., 2023) and DialogCC (Lee et al., 2024) further advanced this line by synthesizing massive image-dialogue pairs using LLM-driven generation pipelines. In contrast, video-grounded dialogue research started with datasets such as AVSD (Alamri et al., 2019), Twitch-FIFA (Pasunuru and Bansal, 2018), and VSTAR (Wang et al., 2023b), where multi-turn dialogues were grounded in short or domain-specific videos. VDAct (Imrattanatrai et al., 2025), building on the VirtualHome2KG videos (Egami et al., 2023), extended this direction to long-horizon, event-driven household scenarios and introduced structured knowledge graphs (KGs) for temporal reasoning. However, despite its richer temporal structure, the dataset remains relatively small and relies on limited manual quality control. This highlights the need for scalable, reproducible LLM-based frameworks for dialogue quality refinement.

Quality Control in Language Resource Construction Ensuring data reliability has long been a central concern in NLP resource construction. Previous studies have shown that human annotators often introduce unintended biases or spurious patterns, allowing models to exploit superficial cues rather than learning the intended task (Gururangan et al., 2018; Belinkov et al., 2019; Zhang et al., 2019). As annotation efforts expand in scale, maintaining consistent data quality becomes increasingly challenging, particularly in crowdsourced settings. In the context of multimodal dialogues, these issues often appear as speculative reasoning, emotional attribution, or stylistic drift that detach conversations from their visual grounding. Our work seeks to mitigate such non-grounded conversational tendencies and enhance the factual alignment between dialogue and visual context.

The Role of LLMs in Data Curation LLMs have recently become integral to the data lifecycle, facilitating not only data generation but also evaluation (Jia et al., 2024), annotation (Gilardi et al., 2023; Wang et al., 2023a), and refinement (Zheng et al., 2023). In particular, LLM-based filtering has shown promise in enhancing dataset reliability by automatically identifying low-quality or noisy examples across diverse tasks. While prior work has used single LLMs for data cleansing (Choi et al., 2024) or complex pipelines for scaled filtering (Henriksson et al., 2025), these methods are vulnerable to idiosyncratic model biases or error propagation. Others have used LLM-driven active learning to allow for human correction (Rouzegar and Makrehchi, 2024). In contrast, our framework is distinguished as a calibrated multi-agent panel. By employing multiple LLMs in parallel and relying on the majority vote, we enhance judgment reliability. A human-in-the-loop calibration phase ensures that our LLM judges are aligned with human intent, making the framework particularly effective for filtering subtle, non-grounded utterances in multimodal datasets.

3. VDAct 2.0

3.1. Revisiting VDAct

VDAct established a long-horizon, event-centric video-grounded dialogue benchmark based on 1,000 scenario videos of everyday home activities such as meal preparation, house arrangement, and leisure. Each video is paired with an event-centric KG and three dialogues that were generated by three different human annotator pairs. This yields a total of 3,000 dialogues and approximately 30,000 QA pairs across diverse question types (*e.g.,* descriptive, temporal, explanatory, and quantitative). The benchmark emphasizes temporal reasoning and multi-turn context over single-turn QA.

While the dialogues were human-generated, human processes can be fallible, especially under high-throughput settings. Since each annotator needed to handle a large number of dialogue generations, this can induce fatigue and habituation, which in turn lead to avoidable mistakes such as speculative statements, “off-video” guessing, shifts in tone (*e.g.,* impoliteness or slang), and overly imaginative phrasing that departs from the grounded scene. These issues accumulate subtly

[1] https://github.com/aistairc/VDAct2.0

Dataset	#Videos	#Dialogues	#QA Pairs	Video Source	Avg. Video Length	Avg. Q Length	Avg. A Length	KG
VisDial	120k (images)	120k	1.2M	-	-	5.1	8.2	✗
Twitch-FIFA	49	15,083	15,083	Soccer match	30 secs	68	6.3	✗
AVSD	11,816	11,816	118,160	Crowdsourced	30 secs	7.9	9.4	✗
VDAct	1,000	3,000	30,095	VirtualHome	248 secs	7.8	10.2	✓
VDAct 2.0	2,975	6,356	63,958	VirtualHome	245 secs	8.4	11.3	✓
– VDAct*	996	2,248	22,567	VirtualHome	247 secs	7.9	8.7	✓

Table 1: Comparison between VDAct 2.0, VDAct, and the existing related datasets. VDAct* refers to the subset of original VDAct data that passed the filtering process and is retained in VDAct 2.0.

Dialogue ID: 0203806
Q: Could you describe to me the man's first action and where he was?
A: He was tidying dirty glasses in the kitchen.
Q: Did the man do this activity for a long time?
A: The whole activity took about a minute.
Q: After the man washed the wine glasses, did he do anything else in the kitchen before moving on to his next activity?
A: I didn't see him wash the glasses, just put them in the sink. After that, he went to the living room to read a book.
Q: Where did the man read the book in the living room?
A: He read it on the sofa.
Q: Where did the man get the book from?
A: From a bookcase in the living room
Q: What did the man do after reading the book?
A: He turned on the TV and went to the kitchen to get a drink.
Q: What did the man get to drink from the kitchen?
A: A bottle of wine.
Q: Where did the man drink the wine?
A: He drank the wine straight from the bottle while sitting on the sofa.
Q: After drinking the wine, what did the man do?
A: He cleaned the TV after drinking.
Q: How did the man clean the television?
A: By wiping it with a towel he got out of the bathroom.

Dialogue ID: 0203807
Q: Where was the man?
A: The man began in the kitchen by getting a sponge from the counter and cleaning around the sink.
Q: Did he turn on the water while cleaning?
A: Yes, after the man wiped the area around the sink he turned the water on for a bit and then wiped around the sink again.
Q: What did he clean next?
A: The man picked up two wine glasses from the table and appears to place them into the sink.
Q: What happened after that?
A: Then the man took two dishes from the counter and placed them in the sink as well after wiping the table with a sponge.
Q: How long did it take the man to do that?
A: It was very brief as the dishes were located next to the sink and he did not wash them after placing them into the sink.
Q: Did the man continue cleaning in the kitchen or leave to another room?
A: The man then walked into the living room.
Q: What did he do there?
A: He watched television.
Q: Where did he watch television from?
A: The man watched the television from the sofa.
Q: Where did he go after watching TV?
A: The man walked to the television to turn it off.
Q: What was the man's final activity?
A: After cleaning the kitchen, watching the television was the man's final activity.

Figure 1: Dialogues for a scenario video (02038).

and can erode overall dialogue quality even when the majority of turns look reasonable in isolation. A straightforward mitigation would be to hire more annotators so that each pair handles fewer dialogues. However, simply hiring additional annotators is not tractable at scale due to cost and coordination overhead. Consequently, VDAct 2.0 expands dialogue coverage while complementing the human annotation process with an LLM-assisted quality control that flags undesirable dialogue content.

3.2. VDAct 2.0: Expanded and Refined

We introduce VDAct 2.0, an enhanced release comprising 6,356 dialogues that are verifiably grounded and well-formed across 2,975 scenario videos. VDAct 2.0 consists of the original VDAct portion (2,248 dialogues over 996 scenario videos) plus 4,108 newly collected dialogues based on 1,979 new videos, preserving compatibility with the original task setup while expanding coverage. Table 1 summarizes the key statistics of VDAct 2.0 and compares them with those of the original VDAct and the other related datasets. The example dialogues in VDAct 2.0 are shown in Figure 1.

To obtain new dialogues, we followed the original VDAct dialogue-creation protocol. Six human annotator pairs produced 6,000 dialogues from 2,000 additional scenario videos depicting distinct household activity sequences. Each dialogue is assigned its ID by concatenating its associated scenario video ID and annotator pair ID, following the original VDAct convention. After collection, we applied an LLM-assisted filtering pipeline as quality control to both the newly generated dialogues and the original VDAct dialogues. The goal is to retain dialogues that are visibly grounded in the video and well-formed by systematically flagging QA turns that exhibit undesirable content. In contrast to labor-intensive manual annotation, we used an automated framework that combines multiple LLMs with predefined triggers to pre-screen QA pairs at scale before identifying low-quality dialogues.

To ensure the reliability of the filtering process, we incorporated a human-in-the-loop calibration, where human annotators labeled a small set of QA pairs for triggers to validate multiple candidate LLMs. Those with the highest agreement with human judgments were selected to perform full-dataset filtering. This model selection ensures that only the best-performing, high-agreement LLMs were deployed for the pipeline.

Trigger	Exceptions	Trigger Example	Exception Example
Speculation: Hedges or guesses about facts not directly observed (*e.g.,* "I think", "maybe", and "probably")	(1) Numeric approximations (*e.g.,* "about [sec/min]" or "a few [sec/min]"), (2) Visual ambiguity to describe what's seen (*e.g.,* "looks like/seems/appears", "some kind of") without causal/mental inference	Q: "Was he making dinner?" / A: "**Maybe** he wanted to eat later."	Q: "How long did it take?" / A: "**About** minute." or A: "It **looks like** he places the cup on the table."
Mind-reading: Attributing unobservable desires or feelings (*e.g.,* "he wants.." or "he's happy/sad")	(1) Neutral intention tied to visible activity (*e.g.,* "preparing to" or "decided to do some research") when not hedged, (2) Instrumental necessity such as "he needs a towel", (3) Mental states "relaxing", "memory", and "sense of control" are allowed since they represent as activities.	Q: "Is he **happy** about the result?" / A: "Yes, he is **happy**."	Q: "What did he do after putting away towels?" / A: "He **is preparing to** brush his teeth.".
Impolite/Slang: Slurs, profanity, insults, derogatory address, dismissive slang (*e.g.,* "dude" when used rudely, name-calling)	(1) Everyday phrasal verbs/mild colloquialisms (*e.g.,* "just sits it down"), (2) Emphatic interjections (*e.g.,* "Yes!", "Oh my", "Well, I see"), and (3) Minor grammar/typos (*e.g.,* "Where's he go")	Q: "Why is he **such a loser**?" / A: "That **dude** is an **idiot**."	Q: "What did he do next?" / A: "He grabs the mug and **sits it down**." / A: "**Oh well**, he closes the door."
Imaginative Phrase: Fanciful analogies, *e.g.,* "like a superhero"	(1) Literal brand/object names (*e.g.,* "TruMoo", "Caracao"), (2) Straightforward comparisons of observable properties without fantasy implication	A: "He **teleports** to the kitchen **like a superhero.**"	A: "He picks up a **TruMoo** milk carton from the fridge."

Table 2: The four triggers as undesirable content for the dataset.

3.3. LLM-assisted Dialogue Filtering

VDAct 2.0 verified human-generated dialogues with an automated quality-control layer. The pipeline operates at the QA-turn level to identify predefined target triggers (*i.e.,* speculation, mind-reading, impoliteness/slang, and imaginative phrasing) and aggregates decisions to the dialogue level. To achieve this, we designed the pipeline with (1) *Human-in-the-loop Calibration for LLM Selection* to calibrate high-agreement LLMs to perform the quality control, (2) *Full-Dataset Labeling and Resolution of Ambiguous Cases* to derive target triggers and short rationales for all QA turns, and (3) *Dialogue Filtering* to accumulate turn-level signals to retain well-formed, video-grounded dialogues.

Target Triggers for Heuristic Screening This section defines the four target triggers that mark a QA turn as undesirable. Table 2 provides their definitions, examples, and exceptions. These triggers were selected based on issues observed in the original VDAct dataset that compromise grounding, clarity, or neutrality. Together, they serve as practical criteria for identifying QA pairs that deviate from visually grounded and stylistically appropriate dialogue. A QA turn is labeled as *no* (undesirable) if either the question or the answer contains any trigger; otherwise, it is labeled *yes*.

The first trigger, ***speculation***, targets phrases that express uncertainty, guesses, or hypothetical reasoning over what is visually grounded in the video. Such phrases introduce assumptions about unseen events that cannot be verified through visual evidence. However, exceptions are allowed for numeric or temporal approximations and mild visual hedging because these convey genuine perceptual uncertainty rather than complete speculation. These exceptions maintain natural language flexibility while preserving grounding. The second trigger, ***mind-reading***, prohibits attributing hidden desires, emotions, or intentions of the agents appearing in the videos. This type of inference cannot be visually confirmed and often reflects personal interpretation rather than observation. Exceptions are made for explicit, action-oriented intentions tied to visible preparation or necessity since these describe observable sequences rather than psychological states. The only mental-state verbs permitted are "relaxing", "memory", and "sense of control", which are described directly as activities rather than inferred emotions. The third trigger, ***impolite/slang***, captures profanity, insults, and other socially inappropriate or non-neutral language. This category was introduced because some dialogues, especially under rapid annotation, showed tone drift, including informal speech or rudeness inconsistent with instructions. However, ordinary phrasal verbs, mild colloquial markers, and minor grammatical errors are excluded from this rule. These exceptions ensure that conversational fluency is not penalized when it does not affect politeness or clarity. The fourth trigger, ***imaginative phrase***, addresses figurative or fantastical expressions that inject unobservable causes or fictional analogies. Such creativity, while linguistically vivid, undermines the dataset's grounding principle by introducing non-literal elements. However, literal object labels are exempt because they refer to items within the visual context rather than imagination.

Human-in-the-loop Calibration for LLM Selection To ensure the reliability of our automated filtering pipeline, we performed a human-in-the-loop calibration step to identify the best-performing, high-agreement LLMs for full-dataset labeling. This

Annotator	Annotator	D	T	D+T
Human A	Human B	0.68	0.60	0.57
Human Panel	Qwen3-235B-A22B	0.64	0.55	0.53
	Qwen3-32B	0.64	0.54	0.52
	GPT-5-nano	0.66	0.52	0.50
	Mistral-Large	0.59	0.50	0.49
	GPT-5	0.63	0.50	0.48
	GPT-4.1	0.62	0.48	0.47
	Deepseek-R1-Distill	0.63	0.47	0.47
	GPT-5-mini	0.60	0.45	0.44
	GPT-4o	0.54	0.43	0.43
	Llama-3.3-70B	0.51	0.38	0.39

Table 3: Cohen's Kappa between individual models and the human panel on 454 QA pairs, ranked by the agreement of the combination of decisions and triggers (D+T) in descending order.

Q-235B	Q-32B	GPT-5n	M-L	GPT-5	D	T	D+T
✓	✓	✓			0.75	0.63	0.62
✓	✓			✓	0.69	0.59	0.58
✓		✓	✓		0.71	0.61	0.58
✓	✓	✓	✓	✓	0.69	0.60	0.57
✓		✓		✓	0.72	0.58	0.57
	✓		✓	✓	0.66	0.60	0.56
	✓	✓	✓		0.67	0.60	0.55
✓			✓	✓	0.67	0.58	0.55
✓	✓		✓		0.65	0.58	0.54
	✓	✓		✓	0.68	0.55	0.53
✓	✓	✓	✓		0.60	0.65	0.52
		✓	✓	✓	0.65	0.54	0.51

Table 4: Cohen's Kappa using 454 QA pairs between a human panel and different combinations of 3 models based on the top 5 highest-correlating LLMs, ranked by the agreement of the combination of decisions and triggers (D+T) in descending order. Q-235B, Q-32B, GPT-5n, M-L, and GPT-5 denote Qwen3-235B-A22B, Qwen3-32B, GPT-5-nano, Mistral-Large and GPT-5, respectively.

process begins with two trained human annotators independently labeling a set of 454 QA turns, randomly sampled from the newly generated dialogues. Annotators followed detailed task guidelines with definitions and examples of all four triggers, using a custom-built user interface. Each QA pair was presented alongside the labeling instructions, and annotators were asked to decide whether the pair was desirable or not. For undesirable pairs, they were required to specify the type(s) of trigger present. If a QA pair was particularly ambiguous or difficult to judge, annotators were allowed to mark it as "NA". For any QA turn where the two annotators disagreed, we introduced a single human adjudicator to review the pair and make a final decision. This adjudicator was selected based on their familiarity and understanding of the task. This adjudication step helps maintain labeling accuracy while reducing the overall annotation workload.

In parallel, multiple candidate LLMs were instructed to annotate the same 454 QA pairs using a prompt that mirrored the human task instructions. This prompt includes descriptions of each trigger, output guidelines, and a few-shot examples. However, unlike human annotators, LLMs were also asked to generate short rationales for each decision and identify the question and/or answer spans that contribute to the decision. We evaluated diverse recent open-source and proprietary LLMs, with varying sizes and costs. The open-source candidates include: DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI, 2025), Llama-3.3-70B (Grattafiori et al., 2024), Qwen3 models (Yang et al., 2025) including Qwen3-32B and Qwen3-235B-A22B, and Mistral-Large (Jiang et al., 2023). Proprietary models from OpenAI include: GPT-5-nano, GPT-5-mini, GPT-5, GPT-4.1, and GPT-4o. For each LLM, we computed agreement with the human labels using Cohen's Kappa, as reported in Table 3. Then, we selected the five highest-agreement models and examined combinations among them to identify an LLM panel that further maximized the agreement via majority voting. To balance computational cost and performance, we chose the best-performing subset of three models, as summarized in Table 4. The final LLM panel used on the full dataset consists of Qwen3-235B-A22B, Qwen3-32B, and GPT-5-nano.

Category / Vote Pattern	Count	%
Consensus (3 vs 0 votes) - 83.14%		
All models voted 'yes'	59,221	65.96%
All models voted 'no'	15,668	17.45%
Majority (2 vs 1 votes) - 16.54%		
Qwen3-235B, Qwen3-32B voted 'yes'	2,500	2.78%
Qwen3-235B, Qwen3-32B voted 'no'	3,723	4.15%
Qwen3-235B, GPT-5-nano voted 'yes'	4,304	4.79%
Qwen3-235B, GPT-5-nano voted 'no'	854	0.95%
Qwen3-32B, GPT-5-nano voted 'yes'	1,804	2.01%
Qwen3-32B, GPT-5-nano voted 'no'	1,710	1.90%
No Majority (1 vs 1 vs 1 or with 'NA')	**288**	**0.32%**

Table 5: Voting distribution across 90,072 QA pairs.

Full-Dataset Labeling and Resolution of Ambiguous Cases After selecting the final LLM panel, we applied the same prompt used during calibration to perform full-dataset labeling. Each QA pair was independently annotated by all three selected LLMs with decisions aggregated via majority voting. The Fleiss' multi-rater agreement (Fleiss, 1971) scores between the three LLMs indicate substantial agreement with 0.69 for decisions and 0.66 for triggers. When considering the combined labels, the value drops slightly to 0.59, reflecting a moderate agreement. Overall, these agreements confirm the reliability of the LLM panel on the task.

Table 5 summarizes the voting distribution across the dataset. Among all QA pairs, a strong majority (83.14%) reached full consensus across three

models. Within this set, 65.96% of QA pairs were labeled as desirable, while 17.18% were labeled as undesirable. An additional 17.45% of QA pairs reached majority agreement, contributing another 9.58% of desirable and 7.96% of undesirable examples. The remaining 0.32% of QA pairs had no clear majority, either due to conflicting decisions (1:1:1 split) or inclusion of ambiguous "NA" responses. Focusing specifically on the undesirable QA pairs, we observed that 1.27% of them lacked majority agreement on the specific triggers, suggesting confusion not only in the overall judgment but also in the reason for rejecting the pair. Altogether, 1.59% of QA pairs (1,434 pairs) were marked as unresolved and required additional review. To assess these ambiguous cases, we randomly sampled 150 QA pairs from this unresolved set and asked human annotators to label them following the same procedure as in the initial calibration phase. We found that these QA pairs were often borderline or heavily subjective, with annotator agreement only slightly above the fair agreement threshold based on combined decision and trigger labeling. Given this level of ambiguity and the small proportion of such cases, we chose to adopt a conservative policy, in which all remaining unresolved pairs were discarded and labeled as undesirable.

Dialogue Filtering To filter the dialogues, we applied a threshold based on the proportion of desirable QA pairs within each dialogue. As expected, lowering this threshold increases the number of dialogues retained but compromises overall quality, while stricter thresholds improve dialogue reliability at the cost of dataset size. Balancing these trade-offs, we adopted a 70% desirability threshold, which retains 6,356 dialogues (71%) of the total. This setting achieves a reasonable balance between quality and dataset size. In practical terms, given that each dialogue contains an average of ten QA turns, this threshold corresponds to at most three undesirable QA pairs per dialogue. Dialogues falling below this threshold are excluded from VDAct 2.0.

3.4. Dataset Analysis

Distribution of Questions and Answers by 3-gram Prefix Figure 2 presents the distribution of the first three words in questions (top) and answers (bottom) across VDAct 2.0. The question distribution is dominated by WH-form openings, particularly "What", "Where", and "How", followed by structured verb-subject combinations, forming binary questions, like "Does he", "Did the", and "Was he". These patterns reflect a strong emphasis on factual, event-grounded inquiries, ranging from identity and object type (*e.g.,* "What kind of") to spatial and procedural reasoning (*e.g.,* "Where did he go?", "How many tasks"). The wide branching into verbs such as *use*, *take*, *get*, and *put* suggests coverage of varied physical actions and temporal events common in household scenarios. Answer prefixes show an even stronger regularity, with subject-initial structures, like "He went to" and "He used the", making up the majority of responses. These support direct answers to event-based queries. Binary responses also feature prominently as they align with the frequency of closed-form questions.

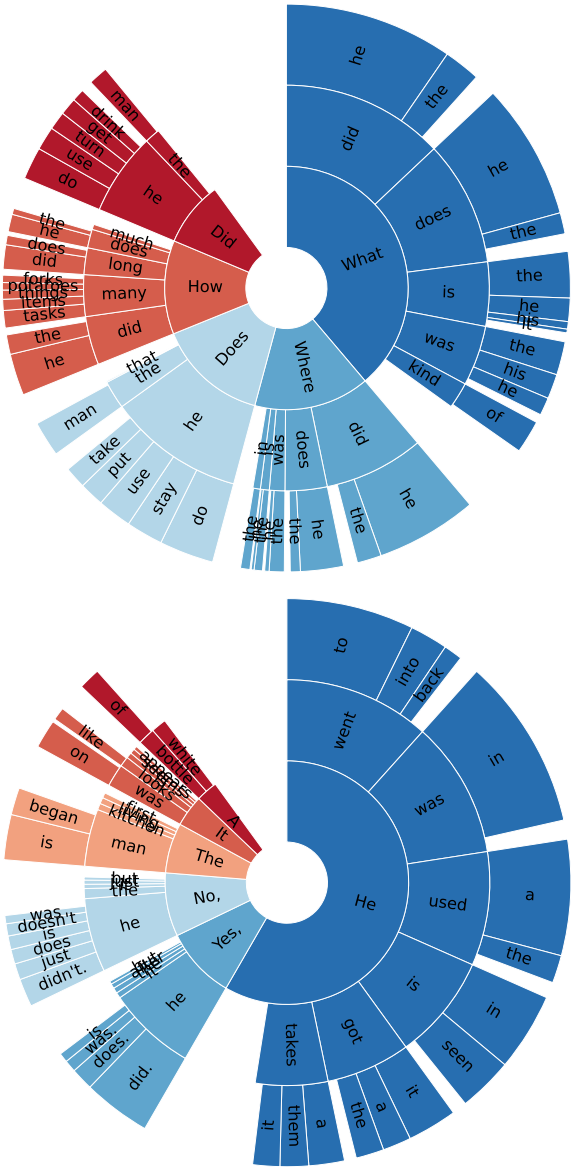

Figure 2: Distribution of *questions* (top) and *answers* (bottom) in VDAct 2.0 based on their first three words.

Dialogue Quality by Annotator Pair Figure 3 shows the percentage of dialogues retained after automated filtering for each annotator pair. The

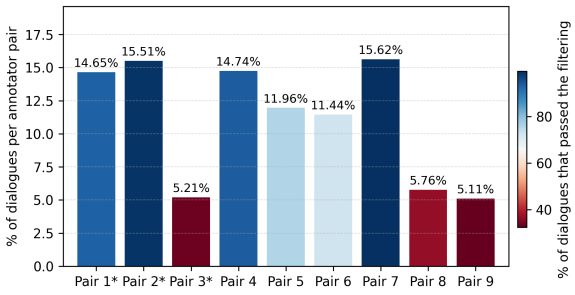

Figure 3: Proportion of dialogues retained in VDAct 2.0 for each annotator pair. Bar colors encode the proportion of desirable dialogues that passed the automated LLM-assisted filtering process in proportion to the number of created dialogues by each pair. The bar marked with an asterisk (*) corresponds to the annotator pair from the original VDAct dataset.

height of each bar reflects the proportion of total dialogues in VDAct 2.0 contributed by that pair, while the color indicates the percentage of their dialogues that passed the LLM-based filtering process.

Overall, quality varied significantly across annotator pairs. The highest retention rates were observed for Pair 7 and Pair 4 for new dialogues, and Pair 1 and Pair 2 for existing dialogues in VDAct, all of whom maintained over 90% dialogue retention. These pairs contributed a large share of the final dataset and demonstrated strong adherence to quality standards. In contrast, Pair 3, Pair 8, and Pair 9 had the lowest retention rates, with 30%-40% of their dialogues passed the filtering. This highlights variation in annotation quality across pairs and underscores the value of LLM-assisted filtering for scalable, systematic quality auditing.

Undesirable QA Pairs Among the full set of 90,526 QA pairs in VDAct 2.0, we identified 20,888 pairs (23.1%) as containing at least one undesirable trigger. Within this subset, the most frequent triggers were mind-reading and speculation, present in 59.25% and 52.96% of the undesirable QA pairs, respectively. Notably, 17.72% of all undesirable pairs were marked with both mind-reading and speculation, indicating a frequent co-occurrence of ungrounded psychological inferences and uncertain guesses. In contrast, the more stylistically driven triggers, like slang/impolite and imaginative phrases, were relatively rare, occurring in 3.56% and 4.93% of cases, respectively.

Closer examination of mind-reading examples reveals recurring patterns such as questions like “How’s the man feeling?”, “What are his intentions?”, or “Does he enjoy...?”, often paired with answers like “The man is feeling...”, “He seems to want...”, or “He decides that...”. These reflect attempts to infer mental states, emotions, or goals that cannot be directly observed. Speculative QA pairs, on the other hand, frequently include hedging language in both questions (*e.g.,* “Do you think he...?”, “Can you tell what...?”) and answers (*e.g.,* “I’m not sure...”, “I think he...”, “I don’t think...”), which indicate uncertainty over what the video represents.

For the less common triggers, slang/impolite examples include casual or humorous language like “Poor fella”, “Party animal!”, or “What’s the dude doing?” in questions, and “a little woopsie”, “The silly man” in answers. Imaginative phrases often invoke fictional or exaggerated imagery, with question examples such as “Oh. Magic!”, “Wow! Teleportation!”, and “poison of choice”, and answers like “the ultimate chef”, “lazy Sunday vibe”, or “Fate brought him”. While rare, these can introduce inconsistency in tone and weaken the factual grounding required for robust video-language reasoning.

4. Experiments

4.1. Experimental Settings

VDAct 2.0 Data Splits VDAct 2.0 is divided into three quality tiers based on the proportion of desirable QA pairs per dialogue: (1) ***Diamond*** (*i.e.,* dialogues with at least 90% desirable QA pairs), (2) ***Gold*** (*i.e.,* dialogues with at least 80% desirable QA pairs), and (3) ***Standard*** (*i.e.,* dialogues with at least 70% desirable QA pairs). Each scenario video in VDAct 2.0 contains between one to three dialogues, all of which share the same activity scenario context. A total of 1,053 scenarios contain exclusively Diamond-quality dialogues, yielding 1,889 dialogues in this top-quality subset. The complete dataset is divided into training, development, and test sets, with priority given to allocating Diamond-quality dialogues to the development and test sets to ensure high evaluation fidelity. Specifically, the training, development, and test sets include 5,456, 300, and 600 dialogues across 2,463, 172, and 340 scenarios, respectively. All splits are disjoint at the scenario level to prevent contextual overlap between training and evaluation.

Baselines We select several recent open-sourced and proprietary VLMs to evaluate on VDAct 2.0. The open-sourced models include: Qwen2.5-VL-7B, MiMo-VL-7B, Gemma-3-12B, InternVL3-8B, InternVL3.5-8B, MiniCPM-V-2.6, and Ovis2-8B. The selected proprietary models include: GPT-5-nano, GPT-5-mini, GPT-5, and Claude Sonnet 4.5. Among these models, we also provide results for their LoRA-finetuned versions of the first four open-source models on the list. For reproducibility, all specific model IDs and hyperparameter settings are reported in the Appendix.

Model	BLEU	ROUGE	METEOR	SPICE	LLM-Acc	LLM-Rel	LAVE	VDEval
Qwen2.5-VL-7B (Bai et al., 2025)	5.9	27.1	29.9	25.7	22.0	32.5	30.6	30.9
MiMo-VL-7B (Team et al., 2025a)	8.1	31.6	31.7	26.1	27.6	37.7	35.1	35.3
Gemma-3-12B (Team et al., 2025b)	3.6	25.6	30.8	23.3	25.1	36.6	35.1	33.8
InternVL3-8B (Zhu et al., 2025)	0.8	31.0	31.7	25.9	25.5	36.4	33.6	34.3
InternVL3.5-8B (Wang et al., 2025)	1.7	**32.4**	32.3	26.8	25.9	36.7	33.5	34.3
MiniCPM-V-2.6 (Yao et al., 2024)	**8.2**	31.7	**33.7**	**27.2**	26.2	36.8	33.9	33.9
Ovis2-8B (Lu et al., 2024)	7.9	30.1	32.3	25.7	26.8	38.3	35.5	34.6
GPT-5-nano	3.8	25.7	25.8	21.5	27.9	38.7	37.5	37.5
GPT-5-mini	2.9	24.4	27.8	21.5	36.2	**47.1**	**46.7**	**45.8**
GPT-5	6.5	30.2	28.1	25.6	**36.9**	45.7	43.9	44.4
Claude Sonnet 4.5	2.0	20.9	28.6	20.3	29.8	42.4	41.8	39.7

Table 6: Pre-trained open-source models and proprietary models' performances on VDAct 2.0.

Model	SPICE	LLM-Acc	LLM-Rel
Qwen2.5-VL-7B	31.8 / 34.8	31.0 / 37.0	40.6 / 45.6
MiMo-VL-7B	36.5 / 38.9	37.4 / 43.3	47.0 / 51.8
Gemma-3-12B	35.4 / 36.9	36.0 / 39.8	45.0 / 48.0
InternVL3-8B	37.4 / **42.0**	40.8 / **49.5**	49.9 / **57.8**

Table 7: Performance of LoRA finetuned models trained on either the original VDAct or the VDAct 2.0 training sets, and evaluated on the VDAct 2.0 test set. Results are reported in x / y format, where x and y denote performances when finetuned on VDAct and VDAct 2.0, respectively.

Evaluation Metrics We evaluate generated text with standard overlap and similarity metrics: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and SPICE (Anderson et al., 2016). We additionally report LLM-based metrics: LLM-Acc and LLM-Rel (Maaz et al., 2024), LAVE (Mañas et al., 2024), and VDAct (Imrattanatrai et al., 2025). LLM-Acc judges binary correctness of the response, while LLM-Rel rates response quality on a 0-5 scale. LAVE assigns a 1-3 score using few-shot instructions. VDEval evaluates the generated answers based on the dialogue sessions. All LLM-based evaluations are instantiated with GPT-4o-mini.

4.2. Results and Discussion

Table 6 presents the performance of various VLMs on our VDAct 2.0, evaluated across traditional text generation metrics and LLM-based evaluation scores. Among open models, MiniCPM-V-2.6 achieves the highest BLEU (8.2), METEOR (33.7), and SPICE (27.2) scores. InternVL3.5-8B obtains the highest ROUGE score (32.4), along with competitive METEOR (32.3) and SPICE (26.8) scores. However, its low BLEU score (1.7) suggests that while its responses are informative, they may diverge from reference phrasing in terms of exact n-gram overlap. For LLM-based evaluations, MiMo-VL-7B leads among open models in LLM-Acc (27.6), LLM-Rel (37.7), and VDEval (35.3), while Ovis2-8B achieves the highest LAVE score (35.5). Despite these advances, proprietary models continue to outperform open models across all LLM-based metrics. GPT-5 achieves the highest LLM-Acc (36.9), while GPT-5-mini outperforms all models in LLM-Rel (47.1), LAVE (46.7), and VDEval (45.8).

Table 7 shows the performance of four LoRA finetuned models on the original VDAct and the extended VDAct 2.0 train sets and evaluated on the VDAct 2.0 test set. Across all models and metrics, fine-tuning on VDAct 2.0 consistently leads to performance gains. InternVL3-8B shows the strongest improvement as it achieves the highest absolute scores, with increases of +4.6 SPICE, +8.7 LLM-Acc, and +7.9 LLM-Rel. Similarly, MiMo-VL-7B also benefits significantly, reaching 38.9 for SPICE, 43.3 for LLM-Acc, and 51.8 for LLM-Rel. Qwen2.5-VL-7B and Gemma-3-12B exhibit modest but consistent gains, although their performances remain behind the other two models. While fine-tuning on VDAct 2.0 leads to consistent improvements over the original VDAct, overall model performance still hovers around 40-50% in terms of response accuracy and relevance on most models. This underscores the persistent challenges in video-grounded dialogue, where models continue to struggle with temporal reasoning, visual grounding, and generating contextually coherent responses. Substantial progress is still needed to achieve robust and reliable multimodal dialogue understanding.

5. Conclusion

We present VDAct 2.0, an extended dataset for video-grounded dialogue with a scalable LLM-assisted filtering pipeline. This process combines human-in-the-loop calibration with a high-agreement LLM panel to identify and retain well-formed, visually grounded QA pairs for multi-turn dialogues, while carefully balancing dialogue quality and dataset scale. Benchmarking results on both pretrained and finetuned models demonstrate clear benefits from training on the filtered dataset, while also highlighting the ongoing challenges in multimodal reasoning and dialogue modeling.

6. Ethics Statement

The dataset includes human-annotated dialogues, with all annotators operating under informed consent and fair compensation through a third-party crowdsourcing company. Annotation guidelines emphasized factuality, respectfulness, and visual grounding. Our LLM-assisted filtering pipeline incorporates human-in-the-loop calibration to reduce bias and ensure consistency, though some subjectivity and residual model bias may persist.

7. Limitations

The taxonomy of triggers is limited to four major categories and does not capture all possible forms of noise. Expanding the taxonomy to cover the more nuanced issues would require a more fine-grained annotation scheme. Moreover, the dataset remains grounded in scripted, domestic scenarios and English-only dialogues, which may limit generalizability to open-domain, multilingual, or real-world video settings. Finally, the definition of "undesirable" content is task-specific and may not align with broader dialogue applications that value creativity or emotional expression.

8. Acknowledgements

We would like to thank all reviewers for their valuable comments on our work. This paper is based on results obtained from: (1) a project, Programs for Bridging the gap between R&D and the IDeal society (society 5.0) and Generating Economic and social value (BRIDGE)/Practical Global Research in the AI × Robotics Services, implemented by the Cabinet Office, Government of Japan, and (2) a project, JPNP25006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). We used ABCI 3.0 provided by AIST and AIST Solutions with support from "ABCI 3.0 Development Acceleration Use."

9. Bibliographical References

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7558–7567.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

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*, pages 65–72.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343, Online. Association for Computational Linguistics.

Juhwan Choi, JungMin Yun, Kyohoon Jin, and YoungBin Kim. 2024. Multi-news+: Cost-efficient dataset cleansing via LLM-based data annotation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15–29, Miami, Florida, USA. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M F Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Shusaku Egami, Takanori Ugai, Mikiko Oono, Koji Kitamura, and Ken Fukuda. 2023. Synthesizing

event-centric knowledge graphs of daily activities using virtual space. *IEEE Access*, 11:23857–23873.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang,

Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Erik Henriksson, Otto Tarkka, and Filip Ginter. 2025. FinerWeb-10BT: Refining web data with LLM-based line-level filtering. In *Proceedings of the*

Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 258–268, Tallinn, Estonia. University of Tartu Library.

Wiradee Imrattanatrai, Masaki Asada, Kimihiro Hasegawa, Zhi-Qi Cheng, Ken Fukuda, and Teruko Mitamura. 2025. A video-grounded dialogue dataset and metric for event-driven activities. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, AAAI-25.

Jinghan Jia, Abi Komma, Timothy Leffel, Xujun Peng, Ajay Nagesh, Tamer Soliman, Aram Galstyan, and Anoop Kumar. 2024. Leveraging LLMs for dialogue quality measurement. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 359–367, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, Jonghwan Hyeon, and Ho-Jin Choi. 2024. DialogCC: An automated pipeline for creating high-quality multi-modal dialogue dataset. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1938–1963, Mexico City, Mexico. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. 2023. Dataperf: benchmarks for data-centric ai development. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, pages 4171–4179.

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*, pages 311–318.

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 125–136, Brussels, Belgium. Association for Computational Linguistics.

Hamidreza Rouzegar and Masoud Makrehchi. 2024. Enhancing text classification through LLM-driven active learning and human annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 98–111, St. Julians, Malta. Association for Computational Linguistics.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng

Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. 2025a. Mimo-vl technical report.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025b. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. 2023b. VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5036–5048, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.*, 57(5).

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Mitigating uncertainty in document classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3126–3136, Minneapolis, Minnesota. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models.

10. Appendices

10.1. Details of Evaluated Models on VDAct 2.0

We select several recent open-sourced and proprietary VLMs to evaluate and set the benchmark results on VDAct 2.0. The open-sourced models include: Qwen2.5-VL-7B (`Qwen2.5-VL-7B-Instruct`), MiMo-VL-7B (`MiMo-VL-7B-RL`), Gemma-3-12B (`gemma-3-12b-it`), InternVL3-8B (`InternVL3-8B`), InternVL3.5-8B (`InternVL3_5-8B`), MiniCPM-V-2.6 (`MiniCPM-V-2_6`), and Ovis2-8B (`Ovis2-8B`). The selected proprietary models include: GPT-5-nano (`gpt-5-nano-2025-08-07`), GPT-5-mini (`gpt-5-mini-2025-08-07`), GPT-5 (`gpt-5-2025-08-07`), and Claude Sonnet 4.5 (`claude-sonnet-4-5-20250929`). All models were finetuned with a global batch size of 128, learning rate of 2e-5, a total of 10 epochs, LoRA adapters (rank 64, scaling factor α=128), and a cosine scheduler with a warm-up ratio of 0.03. The best checkpoint was selected based on the SPICE score on the development set. We used 32 uniformly sampled frames per scenario video for all pre-trained and finetuned models. During inference, the temperature of 0.1 was used. Finetuning tasks were conducted on 8 H200 GPUs, with a random seed of 42, while the model inferences were performed on a single H200 GPU.