



LLAVIDAL : A Large Language Vision Model for Daily Activities of Living

Dominick Reilly¹ Rajat Subhra Chakraborty¹ Arkaprava Sinha¹ Manish Kumar Govind¹
 Pu Wang¹ Francois Bremond^{3,4} Le Xue² Srijan Das¹

¹ UNC Charlotte ² Salesforce AI Research ³ Inria ⁴ Université Côte d’Azur

{dreilly1, sdas24}@charlotte.edu

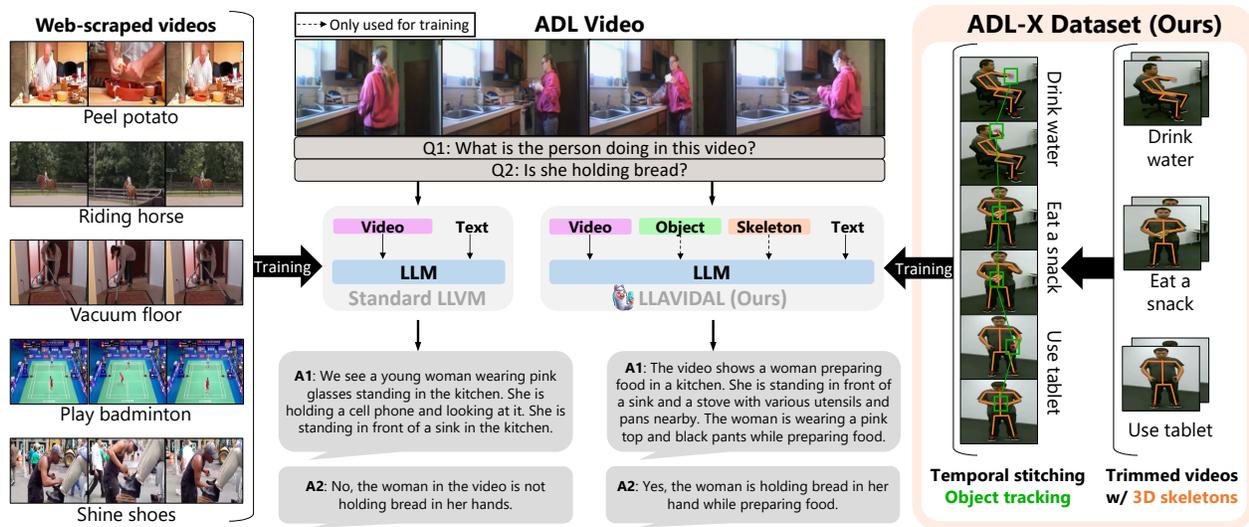


Figure 1. Web-video trained Large Language Vision Models (LLVM) struggle to understand the fine-grained details and human-object interactions present in Activities of Daily Living (ADL). We propose LLAVIDAL, an LLVM trained with **three modalities** on our curated **ADL-X dataset**. ADL-X is derived from trimmed, multi-view ADL videos and is augmented with skeleton and object modalities.

Abstract

Current Large Language Vision Models (LLVMs) trained on web videos perform well in general video understanding but struggle with fine-grained details, complex human-object interactions (HOI), and view-invariant representation learning essential for Activities of Daily Living (ADL). This limitation stems from a lack of specialized ADL video instruction-tuning datasets and insufficient modality integration to capture discriminative action representations. To address this, we propose a semi-automated framework for curating ADL datasets, creating **ADL-X**, a multiview, multimodal RGBS instruction-tuning dataset. Additionally, we introduce **LLAVIDAL**, an LLVM integrating videos, 3D skeletons, and HOIs to model ADL’s complex spatiotemporal relationships. For training LLAVIDAL a simple joint

alignment of all modalities yields suboptimal results; thus, we propose a **Multimodal Progressive (MMPro)** training strategy, incorporating modalities in stages following a curriculum. We also establish **ADL MCQ** and video description benchmarks to assess LLVM performance in ADL tasks. Trained on ADL-X, LLAVIDAL achieves state-of-the-art performance across ADL benchmarks. Code and data will be made publicly available at <https://adl-x.github.io/>.

1. Introduction

This paper aims to explore the potential of Large Language-Vision Language Models (LLVMs) to understand Activities of Daily Living (ADL) which present various challenges including multiple exo-centric viewpoints, fine-grained activities with subtle motion, complex human-object interactions, and long-term temporal relationships. Recently,

[35, 38, 39, 45, 56, 73, 78] have integrated videos into LLMs, leading to the development of video-based LLVMs capable of capturing spatio-temporal features. However, these models are predominantly trained on large-scale web videos [22, 30], which mainly consists of sports clips [8], movie scenes [4, 65], and instructional videos [23, 46, 58]. These videos, typically filmed by professionals, follow strict temporal sequences in closely controlled background (e.g., *Paragliding*). The evident temporal structure and scene semantics in such videos facilitate spatial understanding within LLVMs, as shown in Figure 1. In contrast, ADL videos pose additional challenges, characterized by lack of a strict temporal structure where diverse actions may unfold concurrently within a single sequence [48]. For instance, *a person cooking could intermittently engage in unrelated activities like making a phone call or drinking water, disrupting the linear progression of the composite action cooking*. Consequently, existing LLVMs trained on web videos struggle to capture such visually perplexing dynamics inherent in ADL scenarios. Moreover, unlike specialized video architectures designed for understanding ADL [5, 6, 19, 20, 24, 55, 69], these LLVMs lack explicit utilization of cues like 3D skeletons or human-object interactions (HOIs), which are crucial for understanding ADL. These cues facilitate the learning of view-invariant representations and capture fine-grained details essential for interpreting complex human activities. Hence, the current limitations in understanding ADL stem from the lack of instruction tuning of LLVMs on real-world multiview ADL datasets captured in indoor settings and the simplistic design of LLVMs with holistic operations.

To address these challenges, we first propose a semi-automated framework for curating ADL video instruction tuning data. Unlike existing framework [11, 45], we introduce novel strategies to guide AI annotators in focusing on human-localized spatial regions, incorporating temporal unstructuredness in training distribution, and minimizing hallucinations through weak supervision to generate of high-quality video instruction pairs for training of LLVMs. The result of this framework is **ADL-X** dataset, comprising 100K untrimmed RGB video-instruction pairs, 3D skeletons (S), language descriptions. We then introduce the Large **L**anguage **V**ision model for **D**aily **A**ctivities of **L**iving (**LLAVIDAL**), trained on ADL-X. LLAVIDAL integrates synchronized multimodal inputs from videos, including 3D skeletons and HOI (Human-Object Interaction) cues, into the LLM embedding space. Our study explores multiple strategies for multimodal integration in LLAVIDAL, addressing the challenge of jointly aligning all modalities with the LLM embeddings. To tackle this, we propose a Multimodal Progressive (MMPro) training strategy, enabling effective training of LLAVIDAL. Furthermore, we introduce the **ADL Multiple Choice Question** (ADL

MCQ) and **ADL video description** benchmarks, specifically designed to evaluate LLVM effectiveness in understanding ADL. Empirical results show that LLAVIDAL outperforms other LLVMs, including those trained on datasets ten times larger, on these ADL benchmarks. To summarize our contributions:

- We introduce ADL-X, the first multiview RGBS instruction tuning dataset for ADL tasks, curated via a novel semi-automated framework for training LLVMs.
- We present LLAVIDAL, the first LLVM designed for ADL, integrating 3D skeletons and HOI cues into the LLM embedding space, and propose the *MMPro* training strategy for synchronized multimodal training. LLAVIDAL trained on ADL-X outperforms all baseline LLVMs.
- We establish new benchmarks, ADL MCQ and video description tasks, to assess ADL understanding in LLVMs.

2. Related Works

Towards emulating human cognitive perception in digital intelligence, initial efforts focused on integrating vision and language modalities [26, 49, 54, 62]. Subsequently, the success of LLMs like GPT [7], PALM [14], BLOOM [72] led to the introduction of multimodal conversational models [1, 2, 12, 40, 66, 74, 81] that combine image pixels and LLMs, we call LLVMs. In this section, we briefly discuss methodologies relevant to developing video conversational models.

Image captioners + LLM. With the emergence of foundation models [34, 52, 70, 76], a natural extension is to use pre-trained VLM encoders to map visual inputs into language representations for LLM processing. For instance, CogVLM [68] uses this approach for image captioning, while Socratic Models [75] and VideoChat [35] adapt it for video captioning. Similarly, dialog-based models like VideoChatCaptioner [10] summarize videos via interactions between ChatGPT [7] and a captioner such as BLIP2 [33], while ChatVideo [67] uses task-specific foundation models with ChatGPT [7] to generate responses to user queries. Efficient video processing approaches, which encode video segments or frames via VLMs followed by single or multiple LLM calls with temporal hierarchy, have been extensively explored in [29, 51, 53, 71, 77]. However, these models rely on static image-to-language mappings, lacking explicit modeling of temporal information.

LLVMs. To align image pixels with the LLM embedding space, image-based methods [2, 17, 40, 74, 81] employ a visual connector, often a simple linear projection layer. In contrast, models like Flamingo [1] and BLIP2 [33] use cross-attention mechanisms, such as a Query Transformer (Q-Former), to align visual features with textual prompts. For video alignment, LLVMs similarly use either linear projectors [11, 28, 45] or Q-Formers [35, 56, 78] for feature encoding and alignment. However, these LLVMs generally capture a holistic view without focusing on fine-grained de-

tails or learning view-invariant representations, which are essential for understanding ADLs. In LLVIDAL, additional video-derived modalities enable its visual connector to address these challenges.

More Modalities + LLMs. Methods like VideoLlama [78] and VideoLLaVA [38] incorporate additional modalities such as audio and images alongside videos. Similarly, X-InstructBLIP [50], aligns multiple modalities individually with language embedding space, as in Language-Bind [80]. In contrast, we focus on synchronously aligning multiple modalities, requiring joint rather than separate alignment. This is achieved through our *MMP*ro training strategy, which uses progressive training [32, 60] to sequentially integrate modalities into the LLM embedding space.

3. Background: Training LLM

In this section, we briefly present the general framework for training LLMs. Given an input video $V_i \in \mathbb{R}^{T \times H \times W \times C}$, we encode each frame using the pretrained vision-language model (VLM) CLIP-L/14 [52] to obtain frame-level embeddings $f_i \in \mathbb{R}^{T \times h \times w \times D_v}$, where D_v is the embedding dimension, and $h = H/p$, $w = W/p$ are dimensions adjusted by the patch size p . We extract temporal and spatial features by aggregating these embeddings along their respective dimensions [45]. The resulting video-level features, $\mathcal{X}_i^v \in \mathbb{R}^{F_v \times D_v}$, where F_v denotes the number of spatio-temporal tokens, are obtained by concatenating the aggregated features.

These video features are projected into the LLM embedding space using a parameterized module \mathcal{T}_v , mapping input visual tokens to $Q_v \in \mathbb{R}^{F_v \times K}$. Similarly, the text query is tokenized to $Q_t \in \mathbb{R}^{F_t \times K}$, representing an instructional query from the training data. The inputs to the LLM are then combined as follows [USER: $\langle Q_t \rangle$ Assistant: $\langle Q_v \rangle$]. Optimization of the LLM is performed under a causal language modeling objective: $\min_{\theta} L_{CE}(\text{LLM}(\mathcal{T}_v(\mathcal{X}_i^v)), y_i)$ where L_{CE} denotes the cross-entropy loss, θ are the parameters of \mathcal{T}_v , y_i is the target sequence, and $\text{LLM}(\mathcal{T}_v(\mathcal{X}_i^v))$ is the LLM’s prediction.

4. ADL Video-instruction Pairs (ADL-X)

This section describes the weakly supervised semi-automated data curation framework employed for the creation of a novel dataset, ADL-X. This curation framework is carefully designed to enable the instruction tuning of LLMs within the ADL domain. ADL-X comprises video recordings of ADLs from the NTU RGB+D 120 (NTU) dataset [41]. This selection was motivated by the dataset’s focus on ADL videos and its inherent diversity in terms of actions, subjects, and camera viewpoints.

To facilitate LLM training, question-answer (QA) pairs were systematically generated targeting various aspects of ADL. The semi-automated data curation process for ADL-X incorporates *three* novel techniques as illus-

trated in Figure 2: (i) Person Augmented Generation, which enhances video descriptions with spatially focused human related contents; (ii) Temporal Stitching, which connects short action clips for continuity; and (iii) Weakly Supervised Video Descriptions, which facilitates the generation of high-quality descriptions that guide the creation of video QA pairs. These QA pairs address diverse dimensions of ADLs, including human pose configurations, objects related to human actions, scene appearance, and the fine-grained actions performed. Subsequent sections detail the chronological steps undertaken to develop ADL-X.

Person Augmented Generation (PAG). ADL tasks necessitate focusing spatially on the individual’s postures and their interactions with objects, distinct from the contextual background typical in web videos [19]. Specifically, we utilize skeleton data to crop bounding boxes around individuals. This person augmentation strategy minimizes background information in the video frames, effectively diminishes the generation of extraneous content by LLMs in video descriptions, focusing on human actions.

Temporal Stitching (TS). Real-world ADL videos typically lack temporal structure, in contrast to instructional videos like *cooking*, where actions are sequentially linked. To mimic the inherent randomness of ADLs, we develop a methodology to construct long, untrimmed video sequences by stitching together shorter clips. Initially, we generated 160 composite action sequences using GPT prompts to combine individual actions from the NTU dataset’s 120 actions (A_1, A_2, \dots, A_{120}). An example of such a sequence is *drink water* \rightarrow *eat snack* \rightarrow *wipe face*. The corresponding short clips from the NTU dataset are then stitched together, ensuring that all clips within a sequence maintain the same subject for consistency. To increase diversity and minimize bias towards specific subject-action pairings, we randomize both the sequence of actions and the subject assignments. This procedure yields **16,206 stitched videos**, with a maximum of 7 actions per video.

Weakly Supervised (WS) Video Descriptions. The generation of video descriptions, integral to automating QA pair creation, involves a two-step process: initially, image captions are generated for each frame in the video at a rate of 0.5 fps using CogVLM [68]. Subsequently, these frame-level captions are synthesized into a cohesive video description via one LLM call. However, this step is prone to introducing hallucinations, potentially degrading the dataset’s quality. To mitigate this, we employ weak supervision by incorporating the action sequence obtained from the short clips into the GPT-3.5 Turbo model, guiding it to generate a structured, coherent description limited to 300 words. This method of weak supervision helps eliminate irrelevant noise from the caption generation process. For details on our prompting strategy, see Appendix G.

Generating QA Pairs. To generate domain-specific

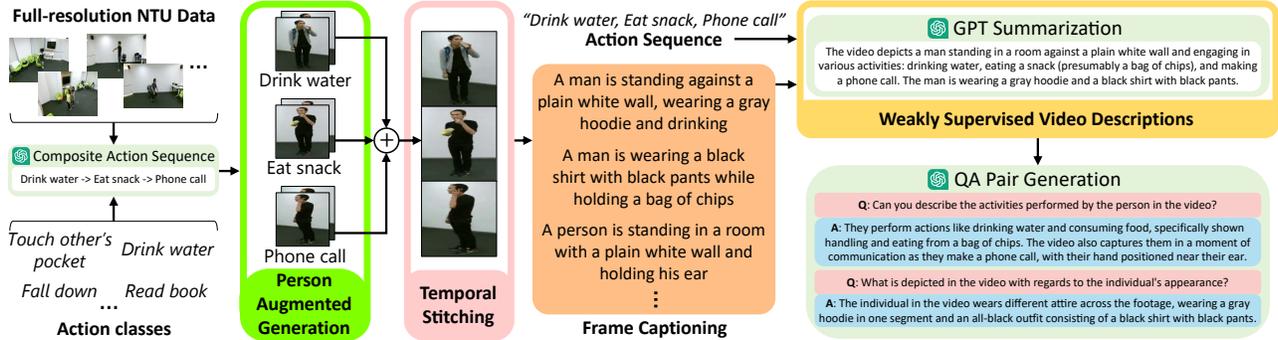


Figure 2. **ADL-X dataset curation pipeline.** The ADL-X dataset is derived from the NTU RGB+D 120 dataset through the use of three techniques: **Person Augmented Generation**, **Temporal Stitching**, and **Weakly Supervised Video Descriptions**. The pipeline leverages CogVLM [68] for frame-level caption generation and GPT-3.5 Turbo [7] for summary synthesis and question-answer pair generation.

QA pairs for ADL, we utilize the dense video descriptions produced in the preceding step. An instruction template (detailed in Appendix G), guides the GPT-3.5 model to formulate questions across various ADL-relevant categories. These categories include video summary, performed actions, spatial details, human-object interactions and other video-specific inquiries. Through this prompting approach, we curated a dataset comprising **100K video instruction pairs**, named ADL-X, derived from the stitched ADL videos. Note that this data curation framework, applied to ADL-X, is adaptable and can be extended to other existing datasets for domain-specific training of LLMs.

5. Multi-modalities of ADL-X

Following the literature for understanding ADL [5, 55], we recognize the importance of integrating additional modalities \mathcal{M}_m , such as 3D skeletons (\mathcal{M}_s) and human-object interactions (\mathcal{M}_o), to enhance the understanding of ADLs. These modalities are critical as ADLs predominantly involve movements of essential body parts or joints, facilitating the learning of view-invariant representations. Moreover, understanding not only the semantics of objects but also their trajectories — integral to the actions performed — is essential for developing fine-grained representations. However, the question remains: how should these modalities be integrated within the LLM of the LLM?

To address this, we investigate several strategies to incorporate skeletons and object tracks into the LLM input space. These include utilizing features extracted from dedicated modality-language encoders, augmenting QA pairs with modality-specific information, and contextualizing inputs to the LLM with modality-specific language as illustrated in Figure 3.

\mathcal{M}_m as features. To integrate new modalities into the LLM input, we map modality-specific data from $\mathcal{M}_m \rightarrow \mathcal{X}_i^m$, which is combined with visual features \mathcal{X}_i^v . It is critical that the features \mathcal{X}_i^m , extracted from skeletons and HOI, are aligned to the language domain to facilitate their integra-

tion with the LLM embeddings, analogous to visual features derived from VLM [31]. We provide a detailed methodology for extracting these modality-specific features \mathcal{X}_i^m .

Skeleton Features (SF) - To extract features from the skeleton data \mathcal{M}_s to be fed as input to the LLM, we employ a skeleton-language model, specifically SkeletonCLIP [63]. SkeletonCLIP is a dual-encoder framework that combines a skeleton backbone [79] and a frozen CLIP text encoder [52]. Initially, the skeleton backbone is pretrained on trimmed NTU clips [59] for skeleton action classification. Subsequently, it is fine-tuned to enhance the alignment between skeleton features and language descriptions of actions using cross-entropy supervision. The resulting skeleton features are denoted as $\mathcal{X}_i^s \in \mathbb{R}^{F_s \times D_s}$, where D_s indicates the dimension of skeleton features. These features are used as input tokens to LLMs.

HOI Features (OF) - Extracting HOI features involves two steps: (i) *Action-conditioned object detection* and (ii) *Object Localization and Tracking*. Both steps leverage off-the-shelf models that are effective without the need for additional training. Given a stitched ADL video composed of a sequence of trimmed video segments (denoted as clip_j), the initial step involves extracting action-conditioned object categories from each clip. This is achieved by uniformly sampling 8 frames from each video and employing a pre-trained BLIP-2 model [33] to generate a list of distinct objects observed in the frames. Subsequently, we refine this list using the ground-truth action labels and GPT-3.5. Specifically, for each clip_j in a stitched video, the corresponding action label and the list of detected objects are input into GPT-3.5, which is prompted to identify the object(s) most relevant to the given action. For example, if the objects *plant, chair, bottle, table* are detected in a video labeled with the action *Drinking*, GPT-3.5 is expected to filter out and select [*bottle*] as the relevant object. Refer to Appendix G for our detailed action-conditioned object detection prompting strategy.

In the second step, we perform the spatial localiza-

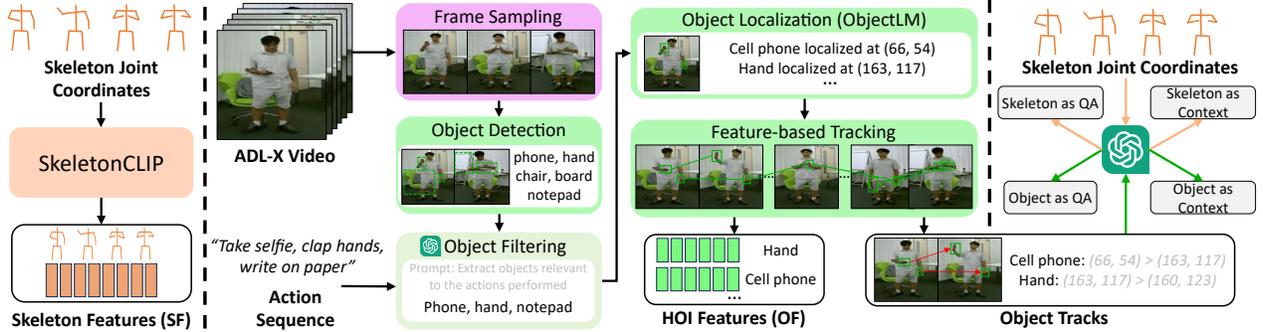


Figure 3. **The multiple modalities of ADL-X.** **Left:** Extraction of skeleton data features (\mathcal{M}_s) using SkeletonCLIP; **Middle:** Pipeline for extracting Human-Object Interaction (HOI) features through action-conditioned object detection, localization, and tracking; **Right:** Outline of our approach for obtaining \mathcal{M}_m as QA and \mathcal{M}_m as context.

tion and temporal evolution (i.e., tracking) of these identified objects within the stitched video. We employ a pre-trained open vocabulary object localization model, **ObjectLM** (OWLv2 [47]), inputting the list of relevant objects along with the corresponding video. For each sampled frame, object bounding boxes are detected, and features for each object are extracted from the image regions within these boxes using ObjectLM. We denote the features for n objects in frame t as $\mathcal{X}_o^t \in \mathbb{R}^{n \times D_o}$, where D_o represents the object feature dimension. To track objects across frames, for each object in frame t , we compute the cosine similarity between its feature vector \mathcal{X}_o^t and all feature vectors in frame $t + 1$ corresponding to the same object category. This object in frame t is then associated with the object in frame $t + 1$ that exhibits the highest similarity score. This matching process is iterated for all objects across each frame, establishing a track for each relevant object throughout the sampled frames. Consequently, for n relevant objects detected across 8 frames, the object features are structured using the following template: $[\langle \mathcal{X}_o \rangle = \langle \mathcal{X}_o^1 \rangle \langle \mathcal{X}_o^2 \rangle \dots \langle \mathcal{X}_o^n \rangle]$ where $\mathcal{X}_o^j \in \mathbb{R}^{8 \times D_o}$ represent the features of each tracked relevant object which are the HOI features in the video.

\mathcal{M}_m as QA. 3D skeleton joint coordinates or relevant object trajectory coordinates are input alongside the associated action sequence into GPT-3.5 Turbo [7], which generates a general description of the skeleton motion or human-object interactions (HOI) of an ADL-X video. This description is then re-fed into GPT-3.5 Turbo to generate two QA pairs that provide detailed explanations of the skeleton and object motions. These QA pairs are then added to the training set of text queries, Q_t , to tune the LLM instruction.

\mathcal{M}_m as context. To integrate contextual information of human skeletons or human-object interactions, we append modality-specific information to the input text query Q_t while training the LLM. For \mathcal{M}_s , we initially identify five peripheral joints – the head, right hand, left hand, right knee, and left knee – due to their significant contribution to motion. For \mathcal{M}_o , we utilize the trajectory coordinates of the relevant object(s) in the videos. Using GPT-3.5 Turbo,

we generate descriptions of the motion for each of these joints or objects based on their trajectories throughout the video, specifically focusing on how the joint and object coordinates evolve. The generated descriptions, denoted as $Q_t^m | m = \{s, o\}$, are subsequently appended to the text query Q_t , incorporating these skeleton or human-object descriptions as additional contextual information. This enriched query $Q_t^{new} = [Q_t^m Q_t]$ is then employed for instruction tuning.

6. Multimodal Progressive (MMPro) Training

We dub our trained LLM on ADL-X as LLAVIDAL. Integrating QA pairs and contextual information into Q_t is achieved using conventional training methodologies for LLMs. However, the joint integration of time-synchronized modalities (\mathcal{X}_v , \mathcal{X}_s , and \mathcal{X}_o) presents challenges, primarily due to conflicting gradients from the different modalities. To address this, we utilize modality-specific connectors that align each modality with the LLM input space. To mitigate the challenges of training with multiple modalities, we adopt a Multimodal Progressive (MMPro) training strategy for LLAVIDAL. This approach incrementally increases the training complexity by progressively adding modality-specific connectors, following a pre-defined growth schedule. These connectors project the modality-specific features into the LLM embedding space, facilitating effective multimodal integration.

MMPro training is structured into $|\eta|$ equispaced stages, with $\frac{\# \text{Total iterations}}{|\eta|}$ iterations per stage. In the case of LLAVIDAL, where we integrate three modalities via connectors into the LLM embedding space, $\eta = 3$ stages. During stage 1, alignment of specific-modality with LLM embedding space is performed. Consequently, video, skeleton, and HOI features are independently projected into the LLM embedding space using linear projection layers \mathcal{T}_m and their respective parameters θ_m for each cue $m = \{v, s, o\}$, resulting in LLM input token representations of the video, skeleton, and HOI cues, respectively:

$$Q_v = \mathcal{T}_v(\mathcal{X}_v; \theta_v); Q_s = \mathcal{T}_s(\mathcal{X}_s; \theta_s); Q_o = \mathcal{T}_o(\mathcal{X}_o; \theta_o) \quad (1)$$

where $Q_m \in \mathbb{R}^{F_m \times K}$. The input to the LLM comprises the concatenation of Q_t and Q_m for $m = \{v, s, o\}$, structured according to the template: [USER: $\langle Q_t \rangle \langle Q_m \rangle$ Assistant:]. This stage 1 training ensures that the video, skeleton, and HOI cues are independently aligned to the LLM embedding space of LLAVIDAL.

Sequence of Integrating Modalities. Our approach to determining the sequence for integrating modalities leverages the principles of curriculum learning. Curriculum learning [21], a machine learning strategy inspired by the pedagogical approach of progressing from easy to complex tasks, is utilized to structure the integration sequence in LLAVIDAL. MMPro training strategy adopts an incremental difficulty progression for multi-modal alignment. The complexity measure guiding the integration sequence is derived from the optimized autoregressive loss obtained from stage 1. Our findings suggest a gradation of integration difficulty among the modalities, with videos and skeletons being relatively simpler to align compared to HOI. The sparsity of HOI features, indicated by the infrequency of distinct HOIs within videos, explains the challenge of aligning HOI with the language embedding space. Consequently, in LLAVIDAL, the modalities are integrated in the order of skeletons followed by HOI, adhering to the curriculum of escalating difficulty.

In the second stage of training, LLAVIDAL’s architecture expands to include additional modality-specific connector. These connector facilitate the simultaneous alignment of video and skeleton data with the LLM embedding. The parameters for this stage, θ_v and θ_s , inherit their initial values from the weights optimized during stage 1. Consequently, the input format to the LLM is structured as follows: [USER: $\langle Q_t \rangle \langle Q_v \rangle \langle Q_s \rangle$ Assistant:] where $\langle Q_t \rangle$, $\langle Q_v \rangle$, and $\langle Q_s \rangle$ represent the text, video, and skeleton query embeddings, respectively. This structured input format ensures a targeted integration of video and skeleton modalities during the MMPro training strategy in stage 2.

The final integration stage in LLAVIDAL incorporates all modalities. The training parameters θ_v and θ_s are further refined from their stage 2 configurations, while θ_o is initialized from stage 1 training. The input to the LLM at this stage includes an additional object modality, formatted as: [USER: $\langle Q_t \rangle \langle Q_v \rangle \langle Q_s \rangle \langle Q_o \rangle$ Assistant:]. This integration approach, as shown in Figure 4, aligns video, object, and skeleton modalities with the LLM embeddings, enhancing the model’s capability to accurately process and understand ADL.

When performing **inference**, LLAVIDAL utilizes only the video cue, consequently eliminating the need for person-centric cropping and additional modalities.

7. Experiments

In this section we first discuss implementation details, then introduce the ADL related tasks used to evaluate state-

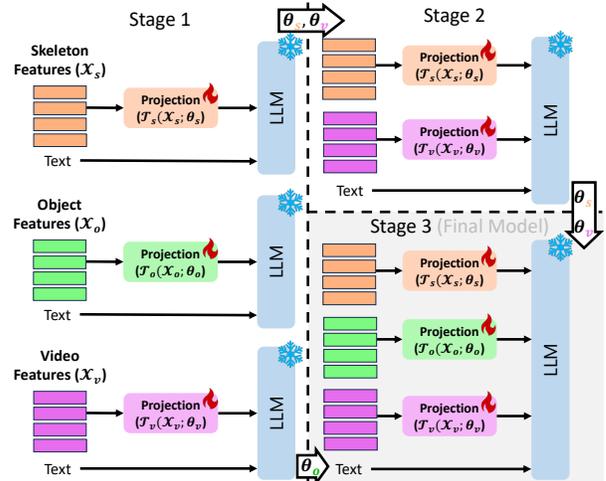


Figure 4. **MMPro training.** Our proposed three-stage progressive training pipeline used to train LLAVIDAL. Stage 1 initializes independent projections for skeleton, object, and video features. Stage 2 combines skeleton and video modalities. Stage 3 integrates all modalities into the final model. Large hollow arrows indicate weight transfer between stages.

of-the-art LLVMs, the evaluation metrics, and finally we present all of the results and analysis.

7.1. Implementation Details

During training, the embedding dimensions are $D_v = 1024$ for visual features extracted from CLIP-L/14 [52], $D_o = 512$ for object features extracted from OWLv2 [47], $D_s = 216$ for skeleton features from SkeletonCLIP [63]. The embedding dimension of the Vicuna [13] LLM is $K = 4096$, and the number of visual and skeleton tokens are set as $F_v = 356$ and $F_s = 256$ respectively. The videos input to LLAVIDAL contain $T = 100$ frames, with a spatial resolution of $H = 224, W = 224$. LLAVIDAL is trained on 8 NVIDIA RTX A6000 GPUs for 3 epochs with a total batch size of 32 and a learning rate of $2e^{-5}$. For the purpose of promoting research in this field, we also provide the extracted skeleton features and object trajectories of LLAVIDAL along with the ADL-X dataset.

7.2. ADL Tasks, Test Datasets, and Metrics

This paper introduces novel benchmarks for assessing LLVM’s temporal understanding of diverse and real-world ADL videos. We propose two new **ADL MCQ** benchmarks including **ADL MCQ-AR** and **ADL MCQ-TC**. **ADL MCQ-AR** involves multiple-choice question-answering for action recognition, where the model selects the correct action from a set of options given a question about the action performed in a video. Similarly, **ADL MCQ-TC** focuses on predicting the missing action in a temporal sequence, given a video containing either a complete or incomplete sequence, including the action at the blank position (e.g., “take something off table”, “drink from cup”, $\langle \text{blank} \rangle$, “walk away”). If the $\langle \text{blank} \rangle$ appears at the end of an incomplete se-

Table 1. Impact of different components in the ADL-X data curation framework on LLM training.

Training data	PAG	TS	WS	ADL MCQ-AR SH	ADL MCQ-TC TSU	Descrip. (Avg) TSU
ActivityNet	✗	✗	✗	39.6	20.9	31.9
NTU120	✓	✗	✓	37.2	20.5	54.1
ADL-X	✗	✓	✓	40.6	25.5	59.1
ADL-X	✓	✓	✗	39.0	24.8	62.2
ADL-X	✓	✓	✓	44.5	29.5	64.8

quence where the input video lacks subsequent actions, the task is an action forecasting. Finally, we also evaluate the LLMs on **ADL video description** tasks for long-duration videos with densely occurring actions and fine-grained captions. Note that all our evaluations are performed zero-shot.

Evaluation Datasets. For ADL MCQ-AR evaluation, we utilize the Charades [61] (Char) and Toyota Smarthome [18] (SH) datasets. Evaluation for ADL MCQ-TC is conducted using LEMMA [27] (LEM) and Toyota Smarthome Untrimmed [16] (TSU) datasets. ADL MCQ-TC on LEMMA is an action forecasting task. Video description tasks are assessed using the Charades and TSU datasets, both featuring long-duration videos with multiple actions per video. Notably, for the TSU dataset, we manually annotated video descriptions with fine-grained details regarding activities performed by elderly individuals, employing 6 human annotators for 174 videos. Our evaluation relies on these annotated descriptions, which we also provide to the community as part of the test set for ADL-X.

Evaluation Metrics. To ensure result consistency and reproducibility, we employ LLAMA 3.1 [66] locally instead of the GPT API, where model weights may update over time. For ADL MCQ task evaluation, we follow [36, 42, 43], prompting LLAMA 3.1 with the output of LLAVIDAL, along with MCQ questions and options, to select the nearest option generated by LLAVIDAL, thus minimizing scoring bias. For video-level description evaluation, following [45], generated descriptions are compared against ground truth and scored by LLAMA across dimensions of Information Correctness, Detail Orientation, Contextual Understanding, Temporal Understanding, and Consistency, with scores scaled to a maximum of 100.

7.3. Impact of ADL-X Training on LLMs

In Table 1, we demonstrate the effectiveness of our novel semi-automated curation of the ADL-X dataset. To validate the importance of domain-specific training, we trained an LLM on trimmed NTU120 clips, showing that the NTU120-trained model outperforms the model trained on ActivityNet [8] for description task, while our ADL-X-trained LLM achieves the highest performance in all ADL tasks. As shown in Table 1, all strategies—**PAG**, **TS**, and **WS**—used to curate ADL-X from NTU120 generate a high-quality video-instruction pairs with minimal hallucinations, enabling the LLM to effectively understand human-centered actions and handle temporal randomness.

Table 2. Strategies to integrate skeletons and HOIs in LLM.

Method	ADL MCQ-AR		ADL MCQ-TC		TSU
	Char	SH	LEM	TSU	Descrip. (Avg)
ADL-X ChatGPT	51.0	44.5	28.6	29.5	64.8
<i>Video+Skeleton</i>					
Skeleton Features (SF)	52.7	42.6	33.1	30.3	66.5
Skeleton QA	47.5	43.1	25.9	29.6	61.8
Skeleton Context (SC)	48.2	45.4	27.8	30.2	66.3
SC + SF	51.2	46.2	33.5	32.5	66.8
<i>Video+HOI</i>					
HOI Features	53.8	48.0	32.6	37.1	68.0
HOI (YOLO+CLIP)	53.7	45.1	34.1	37.0	67.8
HOI QA	50.4	45.5	30.3	28.4	62.5
HOI Context	50.3	44.6	26.5	27.8	63.8

Table 3. Integrating all modalities in LLAVIDAL.

Method	ADL MCQ-AR		ADL MCQ-TC		TSU
	Char	SH	LEM	TSU	Descrip. (Avg)
SF + OF	53.8	40.7	32.1	33.1	65.8
X-InstructBLIP [50]	49.0	45.6	27.5	29.9	65.5
<i>MMPPro Training</i>					
Prog. A (Token)	55.2	48.1	34.3	38.2	70.8
Prog. A (String)	52.7	45.4	32.3	34.5	65.5
Prog. B (Token)	52.8	49.4	32.8	33.0	69.2
Prog. B (String)	51.3	48.6	30.1	32.6	67.4
Prog. A (SF+SC, Token)	54.5	49.3	32.2	34.9	69.5

Verification of ADL-X. To validate the correctness of our LLM-assisted data curation pipeline, we conducted a human evaluation study where the quality of 100 video-QA pairs were rated on a scale of 1-5 based on QA pair correctness. This user study revealed an average rating of 4.1 for the the generated QA pairs, demonstrating that our three key filtering strategies successfully yields high quality video-instruction pairs.

7.4. How to integrate modalities in LLMs?

Table 2 explores the integration of skeleton and HOI cues alongside RGB videos in LLAVIDAL. We integrate skeleton information as features (SF), QA pairs, and context (SC) as illustrated in section 5. Both SF and SC outperform the ADL-X ChatGPT baseline (Video ChatGPT model trained with ADL-X) across most ADL tasks, with the combination of skeleton context and features (SC+SF) achieving the best results, corroborating the importance of skeleton integration for enhanced ADL understanding in LLMs. In contrast, HOI cues used as QA or context provide limited discriminative value for the LLM. However, HOI features extracted from ObjectLM significantly improve performance across ADL tasks, showing their crucial role in ADL understanding. We find that HOI extraction using BLIP2 + OWLv2 outperforms the conventional object detector + CLIP approach, supporting our choice for extracting HOI.

Table 3 shows the performance of LLAVIDAL using various strategies to jointly integrate modalities \mathcal{M}_m , namely videos, skeletons, HOIs, and language. Joint training of all modalities and modality-specific alignment with language following X-InstructBLIP[50] underperforms the

Table 4. **State-of-the-art comparison.** Performance is shown on ADL MCQ and ADL Video Description tasks. [CI: *Correctness of Information*, DO: *Detail Orientation*, CU: *Contextual Understanding*, TU: *Temporal Understanding*, Con: *Consistency*]

Method	Training Data Size	ADL MCQ-AR		ADL MCQ-TC		Charades					Avg	TSU					Avg
		Char	SH	LEM	TSU	CI	DO	CU	TU	Con		CI	DO	CU	TU	Con	
CogVLM [68] + GPT [7]	1.5B Images	52.3	42.5	32.0	23.6	42	62	49.6	36.5	32.8	44.6	55.2	72.0	60.6	30.2	48.5	53.3
CogVLM [68] + Llama [66]	1.5B Images	52.8	43.2	32.5	22.5	40.2	61.8	49.5	36.5	33.5	44.3	49.8	66	56.6	29.8	40.2	48.5
BLIP2 [33] + GPT [7]	1.5B Images	50.2	39.6	28.9	20.2	39.8	60.2	47.8	36.0	37.2	44.2	48.8	66.6	63.6	45.6	39.8	52.9
VideoLlama [78]	2.6M QA Pairs	40.2	44.8	32.6	24.6	22.2	42.5	33.8	20.2	34.5	30.6	57.8	62.0	62.4	48.2	44.4	54.9
VideoLlava [38]	1.2M QA Pairs	41.8	49.2	30.0	25.5	23.6	46.4	34	20.6	33.5	31.6	30.8	54.8	42.4	30.4	44.5	40.6
VideoChatGPT [45]	100K QA Pairs	51.0	39.6	31.4	20.9	26.1	45.2	35.6	21.4	31.2	31.9	31.2	52.8	78.2	64.8	45.6	54.5
ChatUniVi [28]	3M QA Pairs	53.1	48.1	32.3	36.4	36.5	54.5	46.6	32.2	35.9	41.1	56.8	66.9	79.0	50.0	56.6	61.9
ADL-X ChatGPT [45]	100K QA Pairs	51.0	44.5	28.6	29.5	40.6	50.6	49.8	30.6	40.2	42.4	62.4	79.4	70.8	51.2	60.4	64.8
LLAVIDAL	100K QA Pairs	55.2	48.1	34.3	38.2	45.8	64.2	57.0	36.4	39.4	48.6	66.0	86.2	79.6	50.0	72.4	70.8



Q: Describe the video above in details. Include all the actions and object interactions of the person in the scene.

The video depicts a man and a woman in a bedroom. The man is seen bending over a bed, while the woman is standing next to him. The man then proceeds to pull the mattress away from the bed frame, revealing the box spring. He then proceeds to pull the box spring out of the bed frame. The woman then takes the box spring and places it on the floor. The man then pulls the mattress back into the bed frame. The video focuses on the actions of the man and woman in the bedroom, with a particular emphasis on the movements and interactions with the objects in the room.

In the video, a young man is seen working in a bedroom. He is seen laying down a bed, folding a bed, and putting a bed away. The man is wearing a black shirt and a blue jacket. The room has a white floor and a white wall. The man is also seen using a chair and a table. The video is shot in a home in India. The man is seen working alone in the room.

In the video, a person is seen moving a blue blanket or sheet off a bed, possibly to clean or rearrange the bed. The person is also seen walking towards a table with a laptop on it, possibly to use the laptop for work or entertainment.

In the video, a person is seen moving a blue and white blanket from a bed, possibly to clean or rearrange the room. The person is also seen walking towards a bed with a blue pillow, which suggests that they are interacting with the bed and its contents. The room has a white floor, a blue blanket on the bed, and a chair nearby. The person's actions and interactions with the objects in the room, such as the bed and the chair, highlight the ongoing activities within the space.

Video-LLaVA Video-LLaMA Video-ChatGPT LLAVIDAL

Figure 5. **Qualitative results comparing LLAVIDAL with SOTA models.** **Incorrect** descriptions are marked in **red**.

model trained in Table 2 due to difficulties in optimizing the projection layer \mathcal{T}_v for effective alignment of both \mathcal{T}_p and \mathcal{T}_o . In contrast, our *MMPPro* strategy effectively retains discriminative representations from all modalities through its stage-wise integration—first video, then skeletons, and finally HOIs—which addresses optimization challenges faced by baseline multimodal models. To validate the modality integration sequence in LLAVIDAL, we conduct experiments with a curriculum-based model (Prog. A) using stage 1 *MMPPro* losses and an alternative integration sequence—video, then objects, then skeletons (Prog. B). We also ablate the impact of prefixing each modality with a token or string. Results show that Prog. A with token-type prefixing per modality performs best across most ADL tasks, further confirming the efficacy of curriculum-based modality integration in *MMPPro*. Moreover, we find that skeleton features alone are sufficient, and skeleton context is not needed when integrating skeletons in LLAVIDAL.

7.5. Comparison to the state-of-the-art

We compare LLAVIDAL against the state-of-the-art on ADL MCQ tasks and ADL video description tasks.

ADL MCQ. Table 4 compares LLAVIDAL to state-of-the-art LLMs on the ADL MCQ-AR and ADL MCQ-TC benchmarks. LLAVIDAL achieves substantial improvements in action recognition, outperforming the representative baseline, VideoChatGPT, by **+8.2%** and **+21.4%** on the Charades and Smarthome datasets, respectively. On the TSU dataset, LLAVIDAL surpasses VideoChatGPT by up to **+82.7%**, highlighting its superior temporal understanding. Figure 5 provides a visual comparison of LLAVIDAL against representative baselines on the ADL benchmarks, with additional examples in Appendix F.

ADL Video Description. Table 4 shows the performance comparison of baseline LLMs and LLAVIDAL on their video description capabilities on the Charades and TSU datasets. Video-level descriptions are obtained directly from the Charades dataset. For the TSU dataset, comprising lengthy videos, we segment each video into 1-minute clips and input them individually to the LLMs for generating clip-level descriptions. Subsequently, we concatenate all clip-level descriptions and utilize GPT-3.5 turbo to summarize them into a video-level description, following the same instruction template utilized in our dense description pipeline for ADL-X. Across all the 5 VideoChatGPT metrics, the results show that LLAVIDAL consistently surpasses current SOTA methods, and outperforms all models including the image captioners-summarizers pipelines which are trained on billions of images.

8. Conclusion

In this work, we present a novel semi-automated framework for curating ADL video instruction tuning dataset and introduce ADL-X. We introduce LLAVIDAL, an LLM capable of integrating 3D skeletons and HOI modalities using a novel Multimodal Progressive training strategy. To assess LLMs' performance in ADL scenarios, we propose the ADL MCQ and video description benchmarks. Results demonstrate that LLAVIDAL, when trained on ADL-X, surpasses other baselines, indicating its efficacy in grasping intricate temporal relationships within ADL contexts.

Acknowledgments

This work is supported in part by the National Science Foundation (IIS-2245652) and NSF’s NAIRR Pilot initiative (NAIRR240338), which provided credits to access GPT-3.5 Turbo. This work was also supported by funds provided by The University of North Carolina at Charlotte.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736. Curran Associates, Inc., 2022. 2, 14
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 14
- [3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1728–1738, 2021. 13
- [4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 2
- [5] Fabien Baradel, Christian Wolf, and Julien Mille. Human activity recognition with pose-driven attention to rgb. In *The British Machine Vision Conference (BMVC)*, 2018. 2, 4
- [6] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2
- [7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 2, 4, 5, 8, 13
- [8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2, 7, 13
- [9] Marie Chan, Daniel Estève, Christophe Escriba, and Eric Campo. A review of smart homes - present state and future challenges. *Computer Methods and Programs in Biomedicine*, 91(1):55–81, 2008. 24
- [10] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards enriched spatiotemporal descriptions. *ArXiv*, abs/2304.04227, 2023. 2, 13
- [11] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *CVPR*, 2024. 2
- [12] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2, 14
- [13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 6, 15
- [14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 2
- [15] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 15
- [16] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, François Brémond, and Gianpiero Francesca. Toyota smarhome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:2533–2550, 2020. 7
- [17] Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, pages 49250–49267. Curran Associates, Inc., 2023. 2, 14
- [18] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, François Brémond, and Gianpiero Francesca. Toyota smarhome: Real-world activities of daily living. In *Int. Conf. Comput. Vis.*, 2019. 7
- [19] Srijan Das, Saurav Sharma, Rui Dai, François Brémond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *European Conference on Computer Vision*, pages 72–90. Springer, 2020. 2, 3
- [20] Srijan Das, Rui Dai, Di Yang, and François Brémond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2

- [21] Aritra Dutta, Srijan Das, Jacob Nielsen, Rajat Subhra Chakraborty, and Mubarak Shah. Multiview aerial visual recognition (mavrec): Can multi-view improve aerial visual perception? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22678–22690, 2024. 6
- [22] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yanilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thureau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851. IEEE Computer Society, 2017. 2
- [23] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrahm Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kotur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanov, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 2
- [24] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2827–2836, 2015. 2
- [25] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. 14
- [26] Xiaohu Huang, Hao Zhou, Kun Yao, and Kai Han. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. In *International Conference on Learning Representations*, 2024. 2
- [27] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multiview dataset for learning multi-agent multi-view activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7
- [28] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13700–13710, 2024. 2, 8
- [29] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding, 2024. 2, 13
- [30] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2
- [31] Qihong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Learning clip representations for skeleton-based 3d action recognition. *IEEE Transactions on Image Processing*, 27(6):2842–2855, 2018. 4
- [32] Changlin Li, Bohan Zhuang, Guangrun Wang, Xiaodan Liang, Xiaojun Chang, and Yi Yang. Automated progressive learning for efficient training of vision transformers. In *CVPR*, 2022. 3
- [33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 2, 4, 8, 13, 14, 15
- [34] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Jiao Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:12581–12600, 2022. 2, 13
- [35] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2, 13, 14
- [36] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *ArXiv*, abs/2311.17005, 2023. 7
- [37] Fei Liang, Zhidong Su, and Weihua Sheng. Multimodal monitoring of activities of daily living for elderly care. *IEEE Sensors Journal*, 24(7):11459–11471, 2024. 24
- [38] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2, 3, 8, 13, 14
- [39] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 2
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information*

- Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 2, 14
- [41] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3
- [42] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mm-bench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281, 2023. 7
- [43] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 7
- [44] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 13, 14
- [45] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 2, 3, 7, 8, 13, 14
- [46] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 2
- [47] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, pages 72983–73007. Curran Associates, Inc., 2023. 5, 6, 16
- [48] Fatemeh Negin and François Brémond. An unsupervised framework for online spatiotemporal detection of activities of daily living by hierarchical activity models. *Sensors (Basel)*, 19(19):4237, 2019. Published 2019 Sep 29. 2
- [49] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 2
- [50] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning, 2023. 3, 7
- [51] Jong Sung Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S. Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. *ArXiv*, abs/2406.09396, 2024. 2, 13
- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2, 3, 4, 6, 13
- [53] Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S. Ryoo. Understanding long videos in one multi-modal language model pass. *CoRR*, abs/2403.16998, 2024. 2, 13
- [54] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2
- [55] Dominick Reilly and Srijan Das. Just add $\pi!$ pose induced video transformers for understanding activities of daily living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4
- [56] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023. 2, 13, 14
- [57] S. N. Robinovitch, F. Feldman, Y. Yang, R. Schonnop, P. M. Leung, T. Sarraf, J. Sims-Gould, and M. Loughin. Video capture of the circumstances of falls in elderly people residing in long-term care: an observational study. *Lancet*, 381(9860):47–54, 2013. Erratum in: *Lancet*. 2013 Jan 5;381(9860):28. 17
- [58] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhanian, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2
- [59] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 4, 23
- [60] Yi-Ting Shen, Hyungtae Lee, Heesung Kwon, and Shuvra S Bhattacharyya. Progressive transformation learning for leveraging virtual images in training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 835–844, 2023. 3
- [61] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *European Conference on Computer Vision (ECCV)*, 2016. 7
- [62] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022. 2
- [63] Arkaprava Sinha, Dominick Reilly, Francois Bremond, Pu Wang, and Srijan Das. Ski models: Skeleton induced vision-language embeddings for understanding activities of daily living. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 4, 6
- [64] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 15
- [65] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler.

- Movieqa: Understanding stories in movies through question-answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640, 2015. 2
- [66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. 2, 7, 8
- [67] Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *ArXiv*, abs/2304.14407, 2023. 2, 13
- [68] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079, 2023. 2, 3, 4, 8, 13, 15
- [69] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018. 2
- [70] Yi Wang, Kunchang Li, Yizhuo Li, Yanan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2, 13
- [71] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. *ArXiv*, abs/2404.03384, 2024. 2, 13
- [72] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 2
- [73] Antoine Yang, Arsha Nagrai, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. 2
- [74] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2, 14
- [75] Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 13
- [76] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 2
- [77] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *ArXiv*, abs/2312.17235, 2023. 2, 13
- [78] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore, 2023. Association for Computational Linguistics. 2, 3, 8, 14
- [79] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022. 4
- [80] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Liejie Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *ArXiv*, abs/2310.01852, 2023. 3, 14
- [81] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 14



LLAVIDAL : A Large Language Vision Model for Daily Activities of Living

Supplementary Material

A. Overview

The Supplementary material is organized as follows:

- Section B: Related Work
- Section C: Additional Dataset Details
- Section D: Additional Implementation Details
- Section E: Improving Actions: Skeleton Cues vs Object Cues
- Section F: Additional Qualitative Evaluation
- Section G: LLM Prompts Used
- Section H: Limitations
- Section I: Licensing and Intended Use

B. Related Work

In this section, we discuss the recent datasets proposed for instruction tuning of LLMs. We also present the recent advancements in multimodal conversational models both with training-free methods and methods using visual connectors.

Data: Existing video-instruction datasets, such as VideoChat[35], Valley[44], Video-ChatGPT [45], and TimeIT [56], have made significant strides in advancing general video understanding and dialogue. Valley is derived from a website called Jukinmedia that provides videos with diverse categories and wide detailed descriptions. TimeIT dataset from TimeChat offers videos with temporal variations and task diversity. On the other hand, ActivityNet [8] boasts a diverse taxonomy with 203 activity classes, most activity classes are not tailored to the ADL domain. It is to be noted that most LLMs like VideoChatgpt [45], VideoLlava [38] derive their instruction dataset from ActivityNet. Webvid, which is now de-commissioned due to privacy issues introduced in [3], consists of 2.5 million video-text pairs scraped from the web. Although these video instruction datasets are large in scale, offer diverse action classes, and focus on general video understanding, they fail to address the challenges specific to ADL. These challenges include intra-class temporal variations, long-term temporal relationships, complex human-object interactions, and videos captured in multiview settings. Unlike ADL videos, the internet videos in these datasets are predominantly shot by a cameraman, ensuring human-centered frames. Consequently, they do not capture the unstructured randomness spatially and temporally inherent in real-world ADL videos. In contrast, ADL-X is specifically designed to address the challenges inherent in ADL (see Table 5). It captures tem-

poral unstructuredness through the temporal stitching of several unrelated actions in sequence and incorporates complex human-object interactions from the NTU120 dataset. Additionally, our proposed PAG and WS video description techniques effectively eliminate hallucinations, resulting in high-quality video-instruction pairs.

Image captioners + LLM. Advancements in Large Language Models (LLMs) have naturally extended vision-language models [52], leveraging LLMs to enhance reasoning capabilities. The emergence of these foundation models has given rise to training-free methods like Socratic Models [75] and VideoChat [35], which use pretrained vision encoders [34, 70] to map visual information into a language embedding space, followed by LLMs for downstream video tasks.

On the other hand, effective image captioners like CogVLM [68] introduce separate layers into the Transformer block of the LLM to process image features using independent QKV matrices and Feed Forward Networks specifically designed for images. Such effective captioning approaches have inspired methods that map visual information to language via image captioners, followed by processing with LLMs. Among dialog-based models, VideoChat-Captioner [10] summarizes videos through conversations between ChatGPT [7] and a captioner such as BLIP2 [33]. Similarly, ChatVideo [67] employs task-specific foundation models to create a database of "tracklets," with a database manager and ChatGPT [7] collaborating to generate responses for user queries during inference.

For long video understanding, approaches like [71, 77] segment videos into smaller units, providing either segment-level descriptions directly to LLMs or encoding each segment, concatenating tokens, and projecting them into the LLM space. Likewise, Language Repository [29] introduces write-and-read operations to prune text redundancies and extract information across various temporal scales. The Multimodal Video Understanding Framework [53] explores integrating video-specific information into an LLM-based framework by using off-the-shelf vision tools to extract three object-centric modalities from videos and fusing this information through natural language. Additionally, [51] investigates optimal strategies for key-frame selection to significantly reduce redundancies. However, despite these advances, training-free models fail to capture the complex temporal relationships intrinsic to ADL. These

Table 5. Video Instruction Dataset Comparison.

Dataset	Modalities	Subjects	Multiple Views	Videos	QA Pairs	Atomic Actions per Vid	Temporal Rand.	Object Traj.	Type
TimeIT[56]	RGB+L	NA	No	173000	173K	Medium	No	No	Web
VideoChat[35]	RGB+L	NA	No	8196	11K	Low	No	No	Web
Valley[44]	RGB+L	NA	No	64,687	65K	Low	No	No	Web
VideoChatGPT [45]	RGB+L	NA	No	27,801	100K	Medium	No	No	Web
ADL-X	RGB+S+L	106	Yes	16,343	100K	High	Yes	Yes	ADL

relationships, including long-term dependencies and intricate human-object interactions, remain a significant challenge for these approaches.

Large Language Vision Models (LLVMs). The abilities of LLMs in contextual understanding and language generation have led to the introduction of video conversational models. These methods typically employ foundation models to extract visual features from images and project them into an embedding space compatible with language models. Flamingo [1] utilizes a vision-language resampler combined with gated cross-attention, while BLIP2 [33] introduces Q-Former to map image features into the LLM embedding space. Similarly, MiniGPT4 [81] uses a simple linear projection layer. However, these models fall short of becoming conversational assistants due to the lack of human instruction feedback. To address this, mPLUG-OWL [74] first aligns visual and language features through multimodal autoregressive pretraining, followed by multimodal instruction tuning using LoRA [25], enabling more natural and human-aligned responses. Models such as Instruct-BLIP [17] and LLaVA [40] introduce large-scale human instruction datasets to facilitate LLM fine-tuning. Meanwhile, models like PaLI [12] and Qwen-VL [2] allow direct training of LLMs during pretraining or supervised fine-tuning stages.

Other models, including VideoChat, VideoLLaMA, and TimeChat [35, 56, 78], leverage Q-Former for effective feature encoding and alignment. For example, VideoLLaMA [78] employs a vision transformer with an image Q-Former to obtain frame-level representations, followed by a video Q-Former for temporal modeling. Similarly, TimeChat [56] encodes variable-length videos using a timestamp-aware frame encoder with a Q-Former to infuse temporal information into vision tokens, followed by a sliding window Q-Former to condense frame-level features for the projection layer. Building on these approaches, VideoLLaVA [38] jointly trains on both images and videos, pre-aligning visual modalities to language using Language-Bind [80] encoders. VideoChatGPT [45] leverages both temporal and spatial features from videos, obtained by average pooling frame-level features both spatially and temporally.

In contrast to these models, LLAVIDAL incorporates

3D skeleton data and human-object interaction (HOI) cues alongside videos into the LLM embedding space. This integration enables the additional cues to learn discriminative video representations, making LLAVIDAL particularly effective for interpreting ADL videos, where temporal relationships and complex HOIs are crucial.

C. Additional Dataset Details

Question Types. We divide our QA in different questions so that our model understands human object interaction holistically, we lay emphasis on actions performed and the sequence of actions occurring in the video and likewise how objects are associated with the actions. We carefully design such questions relevant to the videos with GPT 3.5 Turbo. The questions encompasses *actions happening, summarization, objects in the scene, color of the objects and questions related to the video*. For Skeleton as QA and Object as QA, we construct two additional questions for each. For Skeleton, we include "What is the motion of the body and joints relative to the actions?" and "Which joints are moving in the video?". For Object, we add "What are the relevant objects in the scene?" and "What is the object in the trajectory $[x1, y1, x2, y2]$ ". These are illustrated in Figures 6 and 7.

Average video and sentence length. There is an average of 23 words per sentence in our QA and average word count for each answer is 42. The average video length is 10 seconds in our dataset. We have 126, 2229 nouns, 551, 172 verbs, 40, 415 actions and 722, 807 objects in our QA showing the overall dynamics of dataset which is illustrated in WordCloud of the Figure 8.

Importance of Person Augmented Generation (PAG). When PAG is not applied to the videos in an ADL dataset, the resulting dense-level captions often include a significant amount of irrelevant information about the background scene. This extraneous information is not directly connected to the subject or the actions being performed, and its presence can introduce noise into the training data. If left unchecked, this noise can have a detrimental effect on the learning process, as the model may erroneously focus on the background details rather than the key elements of the ADL. By failing to isolate the relevant information, the model’s attention is diverted away from the crucial aspects of the task at hand, namely the individual performing the ac-

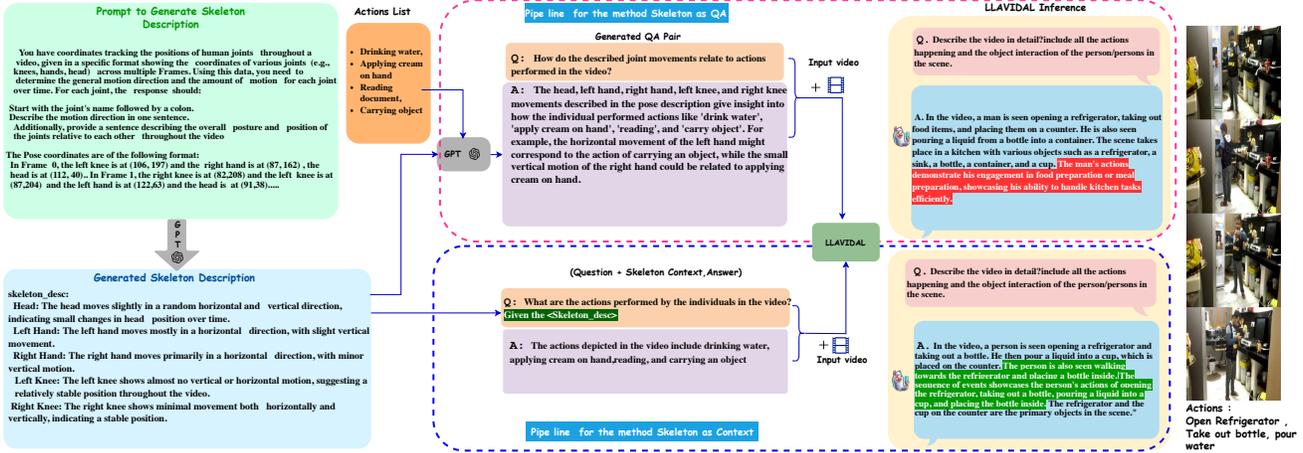


Figure 6. Overview of pipeline for Skeleton as QA and Skeleton as context.

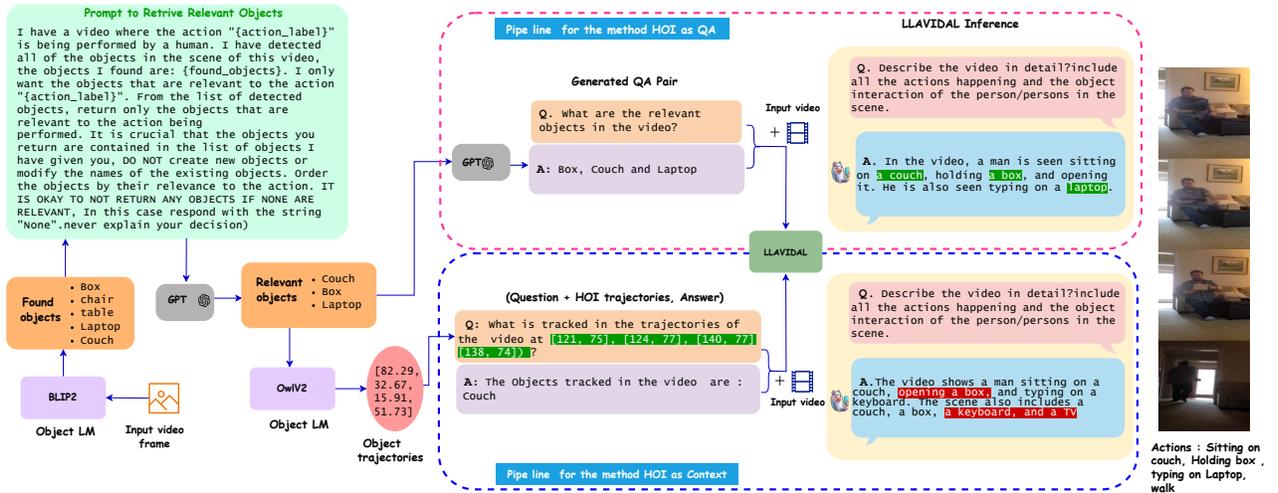


Figure 7. Overview of pipeline for HOI as QA and HOI as context.

tions, the actions themselves, and the interactions between the person and objects in the scene. This dilution of focus can lead to suboptimal performance and hinder the model’s ability to accurately understand and classify ADLs. In contrast, by employing person-centric cropping, the irrelevant background information is effectively eliminated from the videos. This targeted approach ensures that the dense-level captions concentrate solely on the elements that are directly related to the subject and their actions. By maintaining this persistent focus on the relevant information, the training data becomes more coherent and informative, enabling the model to better capture the essential characteristics of the ADLs. In Fig 9, we illustrate an example to highlight the importance of PAG in our semi-automated data curation framework.

D. Additional Implementation Details

We deployed a 4-bit quantized version of CogVLM-17B [68] for annotating frame-level captions. On an A5000 GPU, the inference uses 11GB of memory. The two prompts that are used to get the frame-level descriptions for the ADL-X are – “Give a detailed description of the actions happening and describe the image, include motions and the objects interacted by the person” and “Summarize the content of the image in details explaining all events happening”. CogVLM uses Vicuna v1.5 7b [13] as their large language model and EVA2-CLIP-E [64] as their ViT encoder, the input image dimensions are 224×224 , the average time to annotate a video is 80 seconds at 0.5fps.

LLAVIDAL details. To generate HOI cues, we perform frame-level object detection using BLIP2 and localization using OwlV2. BLIP2 [33] uses a ViT-L and a FlanT5 [15]

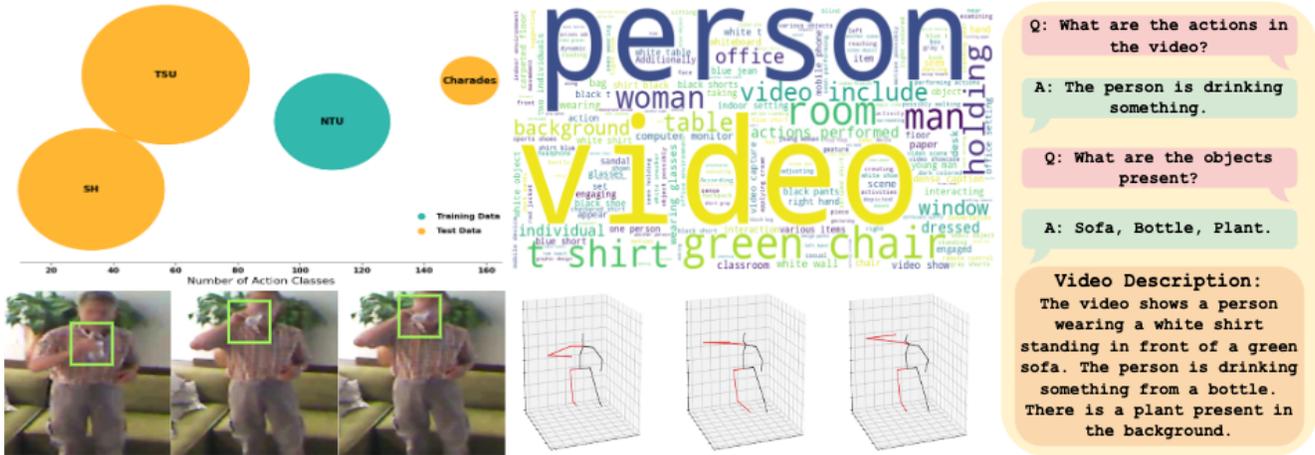


Figure 8. Overview of ADL-X. Top Left: Training and test data distribution; Top Middle: Wordcloud of Textual Representation of Training Data; Bottom Left: Sample video frames with detected relevant object Bottom Middle: 3D skeletons of the corresponding sample video; Right: Sample QA pairs

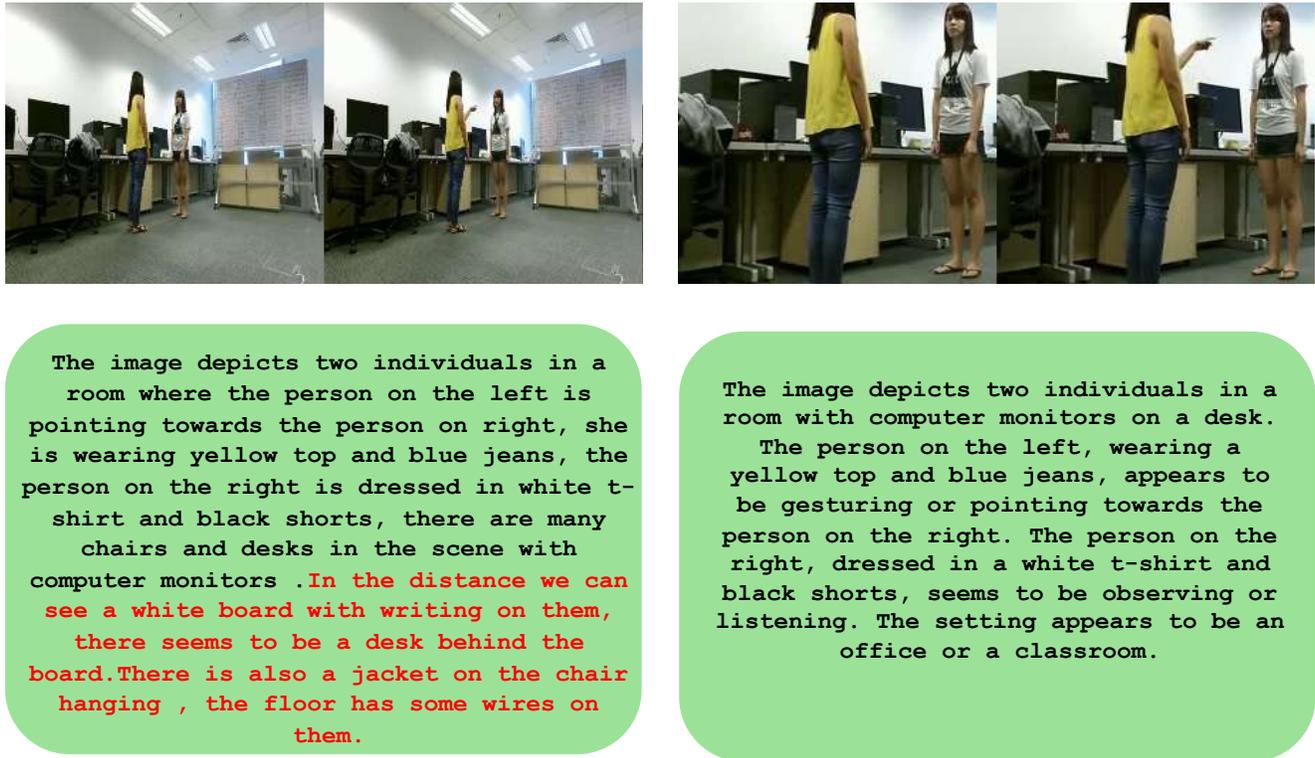


Figure 9. Left: uncropped videos and frame level annotations from CogVLM; Right: PAG and CogVLM captions. The irrelevant information (marked red) adds noise to the annotations.

architecture for detection, while OWLv2 [47] uses an OWL-ViT-L which is a CLIP based model for extracting localization features of the detected objects. In case of skeleton-CLIP, the skeleton Encoder, Hyperformer, is pretrained on NTURGBD for 140 epochs for action recognition, and then is aligned with the CLIP Text Encoder for an additional 100

epochs. LLAVIDAL uses a Vicuna-v1.1 (7B) as the LLM which is frozen during instruction tuning.

E. Improving Actions: Skeleton vs HOI Cues

In this section, we present a comprehensive analysis of our multi-modal approach leveraging HOI tokens, skeleton tokens, and their progressive integration (MMPro) for fine-grained action recognition. Our empirical evaluation across a diverse set of household actions demonstrates distinct performance patterns across modalities, revealing their complementary nature in action understanding when compared against video only trained model. This analysis is shown in Figure 10.

HOI token integration along with video token in LLMs yields substantial performance gains in object-interaction intensive actions, with peak improvements in “use laptop” (+45%), “use tablet” (+32%), “cook cut” (+40%), and “make coffee” (+35%). This performance boost can be attributed to the enhanced object-contextual reasoning capabilities, where the model effectively leverages object-centric features to disambiguate actions occurring in similar spatial contexts. The HOI token cue demonstrates particular efficacy in scenarios requiring fine-grained object state understanding, such as distinguishing between “pour from kettle” (+20%), “pour from can” (+18%), and “pour from bottle” (+15%), where object state transitions and object-specific attributes are crucial for action classification. Additionally, HOI tokens show significant improvements in context dependent actions like “breakfast eat at table” (+25%), “put something on table” (+22%), and “clean dishes dry up” (+18%), where spatial relationships between multiple objects exists.

The skeleton token cue exhibits superior performance in actions characterized by distinctive kinematic patterns, showing significant improvements in body-centric actions such as “take pills” (+28%), “drink from can” (+25%), “drink from bottle” (+22%), and “walk” (+15%). These improvements stem from the model’s ability to capture fine-grained skeletal dynamics, enabling robust discrimination between actions with similar object interactions but distinct motion patterns. Notably, skeleton tokens demonstrate enhanced capability in temporal action modeling, particularly in sequential actions involving multiple body positions such as “lay down” (+18%), “get up” (+16%), and “sit down” (+17%). The skeleton cue also excels in capturing subtle motion differences in drinking actions (“drink from cup” +20%, “drink from can” +25%, “drink from bottle” +22%), where the trajectory and orientation of movement are key discriminative features.

Our proposed MMPro framework, demonstrates significant synergistic effects, particularly in complex actions requiring both object and kinematic understanding. For instance, MMPro achieves significant improvements in “use tablet” (+58%), “make coffee” (+52%), “breakfast take ham” (+45%), and “use laptop” (+48%), where both object state changes and body motion patterns are crucial for

accurate classification. The framework’s effectiveness is particularly evident in ambiguous scenarios where individual modalities underperform, such as “clean dishes” (HOI: +18%, skeleton: +15%, MMPro: +25%), “cook stir” (HOI: +22%, skeleton: +20%, MMPro: +32%), and “pour from kettle” (HOI: +20%, skeleton: +22%, MMPro: +30%).

In complex composite actions like “make coffee” (involving “pour grains” +28%, “pour water” +25%), MMPro successfully captures both the object state transitions and the associated body movements, resulting in more accurate action classification. Our results demonstrate that the MMPro strategy successfully addresses the limitations of video only LLMs, providing a more comprehensive framework for action understanding in complex real-world scenarios.

F. Additional Qualitative Evaluation

In this section, we provide qualitative evaluation of LLAVIDAL and other state-of-the-art LLMs for the tasks of ADL MCQ Action Recognition and ADL MCQ Temporal Completion, illustrated in Figures 12, 13, and 14. In Figure 15, we demonstrate the performance of LLAVIDAL for Video Description Generation on the Charades dataset.

One of the applications of LLAVIDAL is to monitor cognitive decline in geriatric patients through the action forecasting capabilities of our model. In this effort, we have qualitatively evaluated the model on videos of falls on long term care by the IMPL SFU [57]. The subjects in these videos are suffering from dementia, seizure, diabetes like diseases and the dataset contains 175 such falls. We slice the input video before the event of *fall* and prompt LLAVIDAL and other LLM’s to predict whether the person will fall or not. As illustrated in Figure 11, our model outperforms the other LLMs by predicting the fall correctly and by giving proper explanation of why the fall would occur highlighting its reasoning capabilities. While other models predict that the person “has fallen down” and hallucinates the reasoning of the fall as well.

G. LLM Prompts Used

In the following sections, we demonstrate the prompts used:

G.1. Dense Captioning using GPT-3.5 Turbo

```
{“role”:“system”}: "You will play two roles: a human asking questions related to describing a video and an intelligent chatbot designed for video description and dense captioning. Your task is to generate a detailed and descriptive paragraph based on the provided fragmented information about a video."
```

```
“##TASK”：“Users will provide
```

Positive Improvements from Different Modalities vs. Video only model

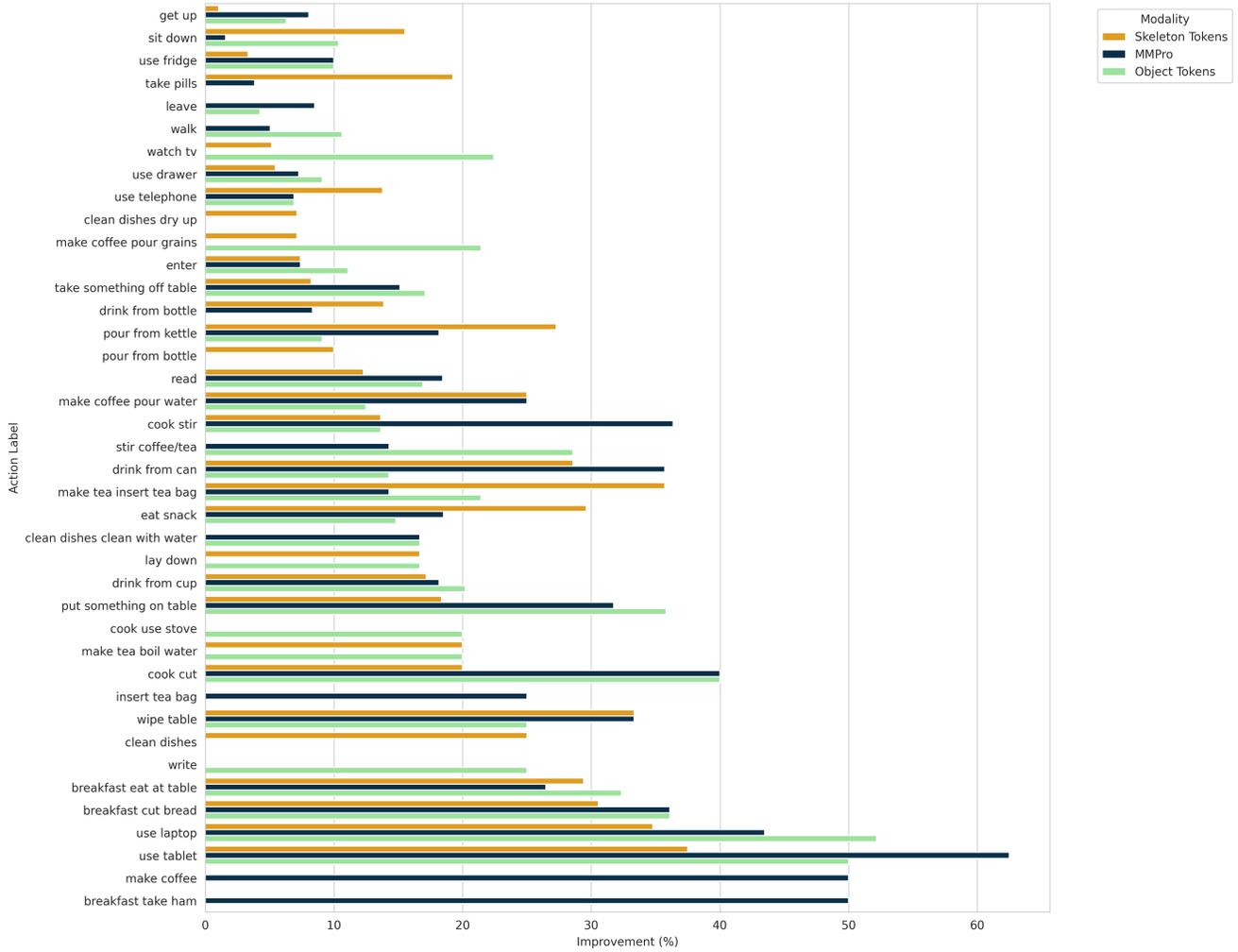


Figure 10. Improvement of actions with object tokens vs skeleton tokens vs MMPro training

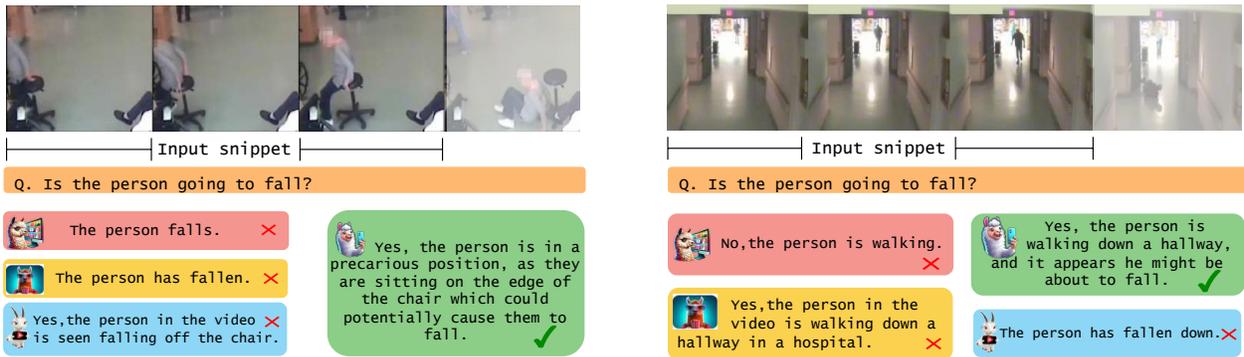


Figure 11. The input snippet is the input video and the grey part is omitted out, here the model needs to detect the greyed action.

fragmented descriptions of a video, and you will generate ONE conversation-like

question and answer related to describing the video in detail. The



Figure 12. Evaluation of ADL MCQ Action recognition task on Charades Dataset

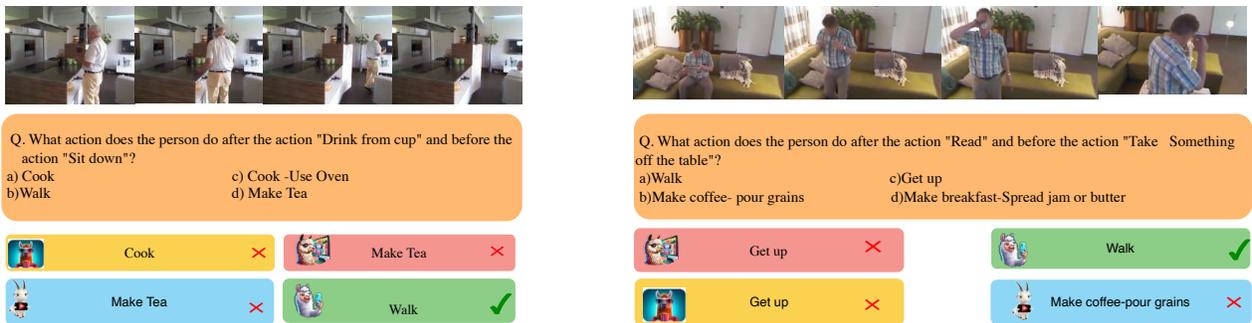


Figure 13. Evaluation of ADL MCQ TC task on TSU dataset

question should ask to describe the video content in detail. The answer should be a paraphrased and well-structured paragraph based on the provided description, with a minimum of 150 words and a maximum of 300 words. When the provided information is short, aim for a 150-word description, and when the provided information is more detailed, aim for very long descriptions up to 300-word description."

###INSTRUCTIONS: "The question must be like a human conversation and focused on describing the video in detail. The answer must be a paraphrased version of the provided information, very detailed and descriptive, and within the specified word count. Combine the information from different sections of the video into a single coherent summary, ignoring any repetitions. Compare the information

across all fragments of video and remove or ignore any inconsistent information and do not say the summary comes from different fragments of the video. Give more emphasis on the actions, the objects, and the colors of the background and the objects. Give the sequence of actions happening in the video and the objects the person interacts with."

{"role": "user"}: "The fragmented video description is: {mega_caption}. Please generate the response in the form of a Python dictionary string with keys "Q" for question and "A" for answer. Each corresponding value should be the question and answer text respectively. For example, your response should look like this: {"Q": "Your question here...", "A": "Your answer here..."}. Emphasize that the answer should focus on describing the video content following the given instructions."

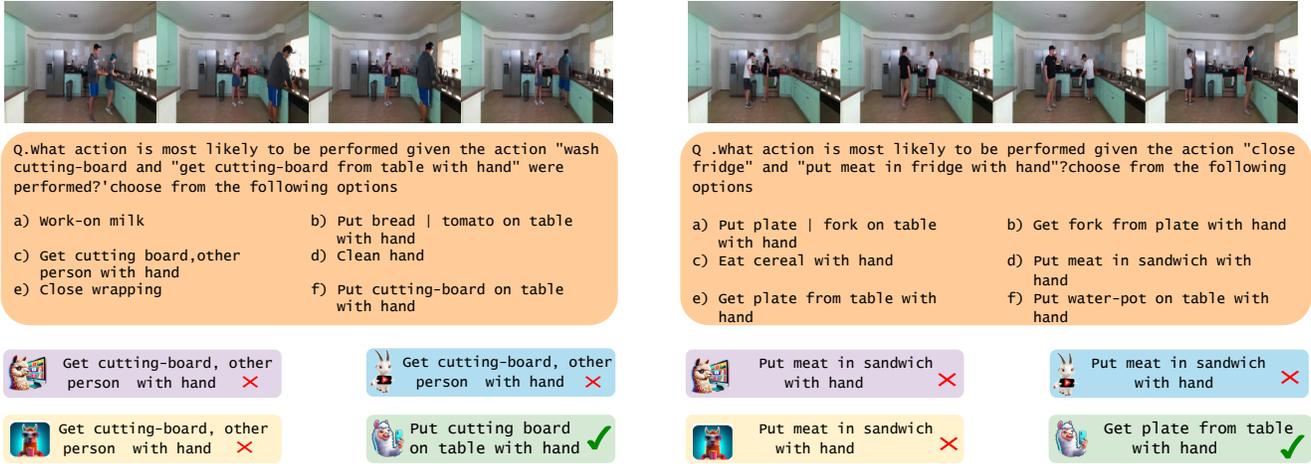


Figure 14. Evaluation of ADL MCQ TC task on Lemma dataset

G.2. QA generation using GPT-3.5 Turbo: Prompt 1

{"role": "system"}: "You play two roles: a human asking questions related to summarizing a video and an intelligent chatbot designed for video summarization and dense captioning. Your task is video summarization. As an AI assistant, assume that you have watched the video and generated the provided caption as the summary of the video. Your task is to play the role of a human who asks three questions related to summarizing the video and then play the role of an AI assistant that provides paraphrased answers based on the video content and the provided caption."

##TASK: "Users will provide a caption of the video alongside dense caption describing detected objects in that scene, and you will generate a set of three conversation-like questions related to summarizing the video. The questions and answers can be very similar, but they should all focus on summarizing the video content. The answers should be paraphrased versions of the provided caption and the dense caption with the object detections. You have information about the video based on the provided caption and have summarized the events in it. You also have the dense caption with the object and scene details. Generate THREE

different questions asking to summarize the video and provide detailed answers to each based on the caption and the dense caption."

##INSTRUCTIONS: "The questions must be like a human conversation and focused on summarizing the video. The answers must be paraphrased versions of the provided caption and the dense caption, and they should be detailed and descriptive."

"-----"

"SAMPLE QUESTIONS:"

"- Can you provide a summary of the video?"

"- What are the main events in the video?"

"- Could you briefly describe the video content?"

{"role": "user"}: "The video caption is: {caption}. The additional dense caption is: {mega-caption}. Generate three different questions on summarizing the video, and provide answers that are paraphrased versions of the given caption and the dense caption. Please attempt to form question and answer pairs based on the two sets of text. Please generate the response in the form of a Python list of dictionary string with keys "Q" for question and "A" for answer. Each corresponding value should be the question and answer text respectively. For example, your response should



Q1. Describe the video in detail.

Q2. Does he drink water ?



A1) The video shows a young boy sitting at a table, writing on a piece of paper. He appears to be focused on his task, and his handwriting is neat and precise. The boy is wearing a casual outfit, with a t-shirt and jeans. The room he is in is well-lit, with natural light coming in from a window in the background. The table he is sitting at is made of wood and has a smooth surface. The boy's writing appears to be in a foreign language, and he seems to be taking his time to ensure that every word is written correctly. Overall, the video captures a peaceful and focused moment of a young boy engaging in a creative activity.

A2) Yes, he takes a sip of water from a glass and then continues to write on the paper.



A1) In the video, we see a young boy sitting at a table in a room. He is wearing a black and white striped shirt. The table is white and has a black and white striped tablecloth on it. There is a white chair in the room, and a black and white striped rug is on the floor. The boy is writing something on a piece of paper. The room is cluttered with various items such as a black and white striped pillow, a white plastic bag, a white bottle, and a black and white striped book. The boy is sitting on a chair, and there is a white box in the room. The video lasts for 30 seconds, and the boy is sitting at the table for the entire duration.

A2) Yes, the young boy in the video is seen drinking water at the table.



A1) The video shows a young boy sitting on a table and writing on a piece of paper. He is wearing a black shirt and appears to be focused on his work.

A2) No, the boy does not drink water.



A1) The video shows a young boy sitting on the floor in a room, wearing a shirt and dark pants. He is writing on a piece of paper, which is placed on a table or a desk. The room appears to be a bedroom, with a white wall. The boy is focused on his writing task, and the room is filled with various items such as clothes, books, and other objects. The overall atmosphere of the scene is calm and peaceful, with the boy engrossed in his writing.

A2) Yes, the boy is drinking water from a glass sitting on a floor.

Figure 15. Qualitative evaluation of video description on Charades Dataset. Text marked in red are incorrect. Text marked in green are correct.

look like this: [{"Q": "Your first question here...", "A": "Your first answer here..."}, {"Q": "Your first question here...", "A": "Your first answer here..."}, {"Q": "Your first question here...", "A": "Your first answer here..."}]. Emphasize that the questions and answers can be very similar, but they should all focus on summarizing the video content."

G.3. QA generation using GPT-3.5 Turbo: Prompt 2

{ "role": "system" }: "You play two roles: a human asking questions related to a video and an intelligent chatbot designed for video summarization and dense captioning. Your task is extracting diverse video information. As an AI assistant, assume that you have watched the video and generated the provided caption as the summary of the video. Your task is to play the role of a human who asks three questions related to summarizing the video and then play the role of an AI assistant that provides paraphrased answers based on the video content and the provided caption."

"###TASK": "Users will provide a caption of the video alongside dense caption describing detected objects, setting and details in that scene, and you will generate a set of three conversation-like questions related to the video. The questions and answers can be very similar, but they should all focus on the details of the video content. The answers should be paraphrased versions of the provided caption and the dense caption with the object and scene details. You have information about the video based on the provided caption and have summarized the actions in it. You also have the dense caption with the scene details. Generate THREE different questions asking the details of the video and provide detailed answers to each based on the caption and the dense caption and one question should be about what actions are happening which should come from captions of the

video."

"###INSTRUCTIONS": "The questions must be like a human conversation and focused on finding the intricate and unique details of the video. The answers must be paraphrased versions of the provided caption and the dense caption, and they should be detailed and descriptive. " "-----"

"SAMPLE QUESTIONS:"

"- What are the actions occurring sequentially in the video?"

"- What are the colors of the outfits of the person in the video?"

"- What are the objects in the scene?"

"- What is the person doing?"

{ "role": "user" }: "The video caption is: {caption}. The additional dense caption is: {mega_caption} Generate three different questions on the details of the video, and provide answers that are paraphrased versions of the given caption and the dense caption. Please attempt to form question and answer pairs based on the two sets of text. Please generate the response in the form of a Python list of dictionary string with keys "Q" for question and "A" for answer. Each corresponding value should be the question and answer text respectively. For example, your response should look like this: [{"Q": "Your first question here...", "A": "Your first answer here..."}, {"Q": "Your first question here...", "A": "Your first answer here..."}, {"Q": "Your first question here...", "A": "Your first answer here..."}]. Emphasize that the questions and answers can be very similar, but they should all focus on the various details of the video content and understanding what actions are happening. Include at least one question about the sequence of actions happening in the video."

G.4. skeleton Description Generation Prompt using GPT-3.5 Turbo

I have the coordinates that track the position of human joints throughout a video. I want to obtain the motion of each of these joints over time, using

only these human joint coordinates. Here are the joint coordinates across observations: {pose_str}. I want to know the general motion of these joints AND the amount of this motion (if the joint moved a lot, or only a small amount over the frames). Respond with a single sentence that INDEPENDENTLY describes the motion directions and amount for each joint over the entire video. Please start your reply for each joint with the name of the joint. What can you tell me about the motion and motion magnitudes of these joints? Describe the concrete direction of the motion of the joints, do not just say they move in many directions, but only describe how it moves and not its numerical coordinates. Do not forget to list the motion and amount of motion in two separate sentences. Begin each description with the name of the joint followed by a colon. Also include a sentence that captures the structure of the human body, such as the posture and position of the joints relative to one another

Here the pose_str, is of the following format:

In observation 0, the right knee is at (104, 201) and the left knee is at (106, 197) and the right hand is at (87, 162) and the left hand is at (134, 49) and the head is at (112, 40). In observation 1, the right knee is at (82, 208) and the left knee is at (87, 204) and the right hand is at (66, 167) and the left hand is at (122, 63) and the head is at (91, 38).....

G.5. Prompt to obtain Relevant Objects using GPT-3.5 Turbo

I have a video where the action "{action_label}" is being performed by a human. I have detected all of the objects in the scene of this video, the objects I found are: {found_objects}. I only want the objects that are relevant to the action "{action_label}". From the list of detected objects, return only the objects that are relevant to the action being performed. It is crucial that the objects you return are contained in the list

of objects I have given you, DO NOT create new objects or modify the names of the existing objects. Order the objects by their relevance to the action. IT IS OKAY TO NOT RETURN ANY OBJECTS IF NONE ARE RELEVANT, In this case respond with the string "None". The relevant objects are (return the objects separated by a comma) (never explain your decision).

H. Limitations

While our approach works well with videos spanning a few seconds, it struggles with long videos. LLAVIDAL's pre-processing pipeline samples 100 frames per video. This sampling rate misses out key information in case of long videos, where there is a larger number of frames. To this end, for the task of generating Video Descriptions, we split the long videos in Toyota Smarthome Untrimmed into clips of 20 seconds each and generate descriptions for each clip. These clip-level descriptions are summarized using GPT3.5 Turbo to obtain a video-level description. However, this summarization step loses valuable information and hence fails to provide an accurate summary of the long video. Future work should explore an effective sampling strategy for long video understanding.

I. Licensing and Intended Use

This paper introduces a large-scale dataset, **ADL-X**, comprising 100K untrimmed RGB video-instruction pairs, 3D skeletons, language descriptions, and action-conditioned object trajectories. The raw videos in ADL-X comprise content from NTURGB+D [59], for which the original authors retain distribution rights for the clipped action videos. The scripts utilized to curate the dataset are open-sourced, facilitating the regeneration of the dataset. We will also provide comprehensive features, including image features extracted using CLIP, skeleton features derived from skeletonCLIP, and HOI features obtained through ObjectLM. We plan to release ADL-X via an academic website for research, academic, and commercial use. The dataset is protected under the **CC-BY** license of Creative Commons, which allows users to distribute, remix, adapt, and build upon the material in any medium or format, as long as the creator is attributed. The license allows ADL-X for commercial use. As the authors of this manuscript and collectors of this dataset, we reserve the right to distribute the data. Additionally, we provide the code, data, and instructions needed to reproduce the main experimental baseline results, and the statistics pertinent to the dataset. We specify all the training details (e.g., data splits, hyperparameters, model-specific implementation details, compute resources

used, etc.). Furthermore, we release the code and model weights of our proposed **Large L**anguage **V**ision model for **D**aily **A**ctivities of **L**iving (**LLAVIDAL**), along with the features and instruction QA pairs for the combination videos. The ADL-X dataset focuses on ADL and does not contain any personal data that can resemble evidence, reveal identification, or show offensive content.

The ADL-X dataset can be used by multiple domain experts to advance research and development in various applications related to ADL. Its potential applications include, but are not limited to, assistive technologies, healthcare monitoring systems [37], smart homes [9], robotics for assisted living, and instructional videos for ADL training and support. The dataset can also contribute to the development of AI-driven solutions that aim to improve the quality of life for individuals with disabilities, older adults, and those in need of daily assistance. While we believe that the ADL-X dataset has the potential to make a positive impact on society by enabling the development of technologies that support and enhance the lives of individuals, we acknowledge that, as with any technology, there is a possibility that the dataset or the ideas it presents could be misused or adapted for harmful purposes. However, as authors, we strongly oppose any detrimental usage of this dataset, regardless of whether it is by an individual or an organization, under profit or non-profit motivations. We pledge not to support any endeavors that could cause harm to individuals or society in relation to our data or the ideas presented herein. Our intention is to foster research and innovation in the field of ADL analysis and support, ultimately contributing to the development of technologies that improve the quality of life for those who need assistance with daily activities. We encourage all users of the ADL-X dataset to adhere to the highest ethical standards and to prioritize the well-being of individuals and society in their research and development efforts.