ABDUCTIVEMLLM: Boosting Visual Abductive Reasoning Within MLLMs

Boyu Chang[1], Qi Wang[1,2], Xi Guo[1], Zhixiong Nan[3], Yazhou Yao[4], Tianfei Zhou[1,5]*

[1] Beijing Institute of Technology
[2] Beijing Institute of Technology, Zhuhai
[3] Chongqing University
[4] Nanjing University of Science and Technology
[5] State Key Laboratory of Environment Characteristics and Effects for Near-space

Abstract

Visual abductive reasoning (VAR) is a challenging task that requires AI systems to infer the most likely explanation for incomplete visual observations. While recent MLLMs develop strong general-purpose multimodal reasoning capabilities, they fall short in abductive inference, as compared to human beings. To bridge this gap, we draw inspiration from the interplay between verbal and pictorial abduction in human cognition, and propose to strengthen abduction of MLLMs by mimicking such dual-mode behavior. Concretely, we introduce **AbductiveMLLM** comprising of two synergistic components: REASONER and IMAGINER. The REASONER operates in the verbal domain. It first explores a broad space of possible explanations using a blind LLM and then prunes visually incongruent hypotheses based on cross-modal causal alignment. The remaining hypotheses are introduced into the MLLM as targeted priors, steering its reasoning toward causally coherent explanations. The IMAGINER, on the other hand, further guides MLLMs by emulating human-like pictorial thinking. It conditions a text-to-image diffusion model on both the input video and the REASONER 's output embeddings to "imagine" plausible visual scenes that correspond to verbal explanation, thereby enriching MLLMs' contextual grounding. The two components are trained jointly in an end-to-end manner. Experiments on standard VAR benchmarks show that **AbductiveMLLM** achieves state-of-the-art performance, consistently outperforming traditional solutions and advanced MLLMs.

Code — https://github.com/ChangPtR/AbdMLLM.git

Introduction

Visual abductive reasoning is the process of forming an explanatory hypothesis for incomplete visual observations. It is an integral part of human cognition (Peirce 1931; Shanahan 2005) and humans routinely employ it in everyday life, both *verbally* and *pictorially* (Thagard and Shelley 1997). Given the observation O: `'the street is wet and the roof is dry'`, one might verbally abduce that a water truck has recently passed by and sprayed the street, based on some hidden governing rules such as `'rain wets both streets and roofs'` and `'a water truck wets only the street'`. Alternatively, one might pictorially abduce by forming a mental picture of a water truck spraying water as it driving down the street. This imagined scene resembles the hypothesized event in a more direct and concrete way than a verbal or sentential representation would, and can in turn facilitates verbal abduction. This very ability gives humans a distinct advantage over machines in high-level reasoning, and represents one of the most valuable capacities to be emulated in modern machine vision system.

Recently, multimodal large language models (MLLMs) have emerged as promising foundations for building visual reasoning systems (Wu et al. 2023). Trained on vast amounts of human knowledge, these models have developed impressive capabilities in multimodal reasoning tasks such as visual question answering, multimodal dialogue, and visual chart reasoning. However, recent studies (Wang et al. 2024b; Chinchure et al. 2024) have highlighted a significant gap between current MLLMs and human abduction capability in understanding ambiguous observations.

To address this limitation, we draw inspiration from human cognition, where verbal and pictorial abduction interact to interpret incomplete visual observations. Based on this insight, we introduce **AbductiveMLLM**, which enhances VAR capabilities of MLLMs by integrating complementary verbal and pictorial abductive process. Specifically, our method consists of two synergistic components: 1) a REASONER, which extracts and selects high-quality verbal hypotheses from an LLM, serving as targeted priors for MLLMs to generate plausible explanations; 2) a IMAGINER, which simulates the pictorial thinking process using diffusion models to guide and refine MLLM-generated explanation. These two components are jointly optimized in an end-to-end manner to allow interactions between verbal and pictorial modes of abductive thinking, and narrows the gap between abstract reasoning and concrete imagination in a more similar way as human cognition.

More specifically, REASONER begins by prompting a LLM to generate a diverse set of candidate hypotheses based solely on video captions. While these hypotheses incorporate broad commonsense knowledge, they often lack grounding in the actual visual content and may therefore be causally inaccurate. To address this, we introduce a *causality-aware contrastive learning* mechanism that pro-

*Corresponding author

motes alignment between the observed video and causally relevant hypotheses (rather than relying on superficial similarity). This filtering process effectively prune out spurious candidates based on causal relevance and narrows the reasoning search space. The filtered hypotheses are then passed to an MLLM to generate a verbal explanation. To complement verbal reasoning with visual imagination, IMAGINER takes both the observed videos and output embeddings from the MLLM as conditioning signals for a generation model. Rather than training a new video generator from scratch, we adapt an existing text-to-image diffusion model (*i.e.*, Stable Diffusion (Rombach et al. 2022)) with lightweight spatiotemporal adapters. As in (Wang et al. 2025b; Ma et al. 2025), the generator is not used to produce high-quality visual results, but instead serves as a reasoning guide: a latent denoising loss is applied to encourage the model to converge on visually plausible outcomes.

Contributions. This work presents **`AbductiveMLLM`**, which represents a pioneering effort in enhancing the abductive capability of MLLMs. • From verbal perspective, we develop a causality-aware contrastive learning model to mine high-quality textual hypotheses, reducing reasoning space and providing crucial priors for MLLMs. • From pictorial perspective, to the best of our knowledge, this is the first study to visual abductive reasoning that explicitly incorporates pictorial thinking capability to improve verbal abductions, inspired by human cognitive processes. Our method shows promising performance on standard benchmarks, consistently outperforming existing specialized small-scale models and MLLMs, setting the new state-of-the-art.

Related Work

Visual Abductive Reasoning (VAR). VAR aims to infer the most likely explanation for partially observed visual events. Early VAR approaches addressed only static images. (Hessel et al. 2022) introduced the Sherlock dataset and adapted CLIP for image-based abductive inference. RCA (Zhang, Ee, and Fernando 2024) augmented this by a visually guided multi-head attention mechanism and a revised contrastive loss. BiGED (Tan et al. 2025a) proposed a relational GNN to predict human pre-action sequences in indoor scenes from a single image. However, these approaches are constrained by the static and incomplete nature of single-frame observations, and often fail to capture the complex spatiotemporal causal structure of open-world scenarios.

To address the limitations above, recent works have shifted toward video-based abductive reasoning. REASONER (Liang et al. 2022) is among the first to build a dataset of real-world visual event sequences, and combines a causality-aware video encoder with a cascade decoder, enabling abductive reasoning over arbitrary visual events. Subsequently, UPD-Trans (Xu et al. 2024) introduced probabilistic distillation in a Transformer. Conan (Xu et al. 2023b) builds an agent to explore active interaction in simulated environments. Videoabc (Zhao et al. 2022) applies hierarchical reasoning to capture long-term dependencies. Knowledge integration has also been explored: KN-VLM (Tan et al. 2025b) introduces visual knowledge from observed videos and textual knowledge from an external knowledge base. MAR (Li et al. 2023) and COIN (Li et al. 2024b) incorporate symbolic reasoning to enhance abduction of human actions. All these studies are specialized small-scale models and focus exclusively on verbal reasoning. In light of recent advances in MLLMs, we pivot to enhance MLLMs for visual abductive reasoning. Drawing on human cognitive processes, we propose a unified multimodal network for video-based abductive reasoning, which integrates both verbal and pictorial thinking.

Multimodal Large Language Models (MLLMs). MLLMs have emerged as leading paradigm for video understanding. Mainstream approaches typically build upon pre-trained large language models, integrate dedicated video encoders, and employ techniques such as self-supervised learning and instruction tuning to achieve effective vision-language alignment and enhance multimodal representation capabilities (Wang et al. 2024c; Lin et al. 2024; Lyu et al. 2023; Chen et al. 2023; Zohar et al. 2024; Wang et al. 2024a). These models have been successfully applied to a wide range of multimodal tasks, including video question answering (Wang et al. 2024d; Maaz et al. 2024; Wang et al. 2025a), multimodal dialogue (Luo et al. 2023), video captioning (Cheng et al. 2024; Xu et al. 2023a), *etc.*. Nevertheless, recent studies (Wang et al. 2024b; Chinchure et al. 2024) have underscored a significant gap between current MLLMs and human abductive capability, which indicates that these models still lack the capacity for advanced reasoning grounded in causal relationships. Our work proposes enhancements from verbal and pictorial modalities, and narrows the abductive reasoning capability between MLLMs and humans.

Our Approach

VAR Task. We follow the task definition of VAR in (Liang et al. 2022). Given a video sequence containing T events $\mathcal{V} = \{O_1, \ldots, O_{t-1}, H, O_t, \ldots, O_{T-1}\}$, where the events are logically related and chronologically organized. Among them, $\mathcal{O} = \{O_t\}_{t=1}^{T-1}$ denotes the collection of T–1 observed premise events, and H represents unobserved explanatory event. Notably, H may occur at any position within $\mathcal{V}$. The goal of the VAR task is to infer the most likely verbal explanation E_h for the unobserved event H, based on the observed events in $\mathcal{O}$.

Main Idea. Inspired by how humans integrate verbal and pictorial thinking for abductive reasoning, we introduce a joint network to enhance the abductive reasoning capabilities of MLLMs. **`AbductiveMLLM`** consists of two main modules: REASONER and IMAGINER. REASONER first generates candidate hypotheses with a blind LLM, then selects causally relevant hypotheses through cross-modal causal contrastive learning, which enhances MLLMs' reasoning in verbal mode. IMAGINER is a diffusion model with lightweight adapters, conditioning on REASONER output embeddings and observations. It is trained end-to-end with REASONER to provide enhancement in pictorial mode. Fig. 1 illustrates the entire process of our method.

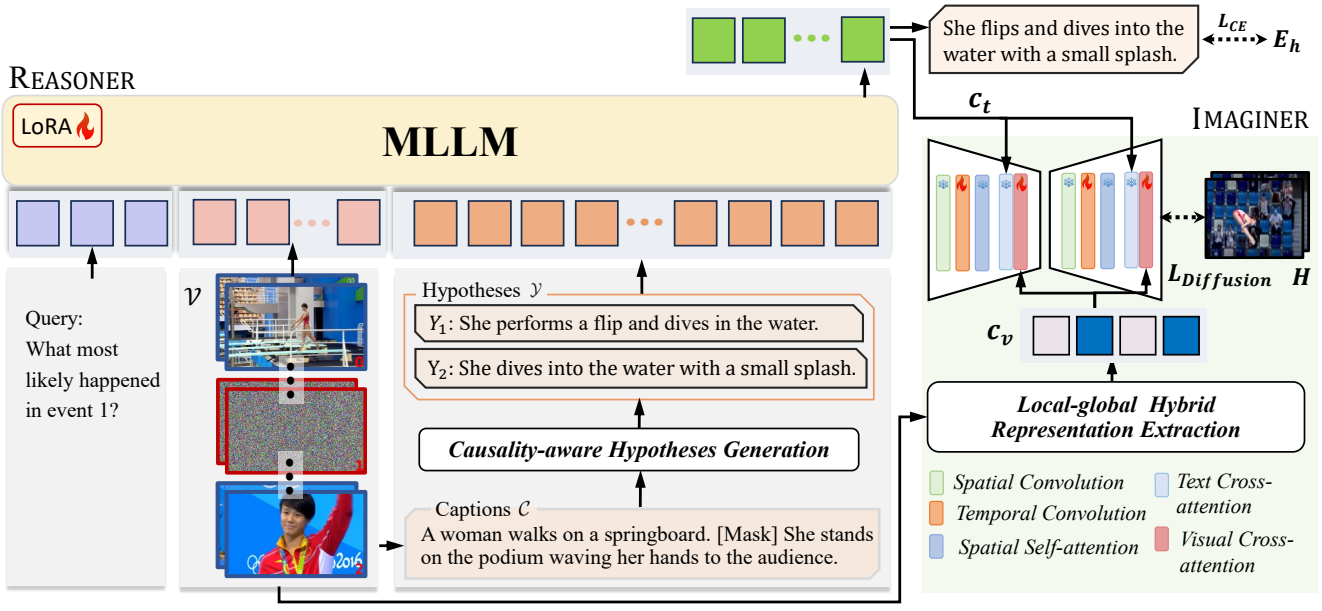

Figure 1: **Network architecture of `AbductiveMLLM`.** The network consists two synergistic components: REASONER and IMAGINER. On the left, REASONER takes a query and a sequence of incomplete observations $\mathcal{V}$ as input. First, it generates captions $\mathcal{C}$ for each observed event. Based on $\mathcal{C}$, Causality-aware Hypotheses Generation module provides high-quality hypotheses $\mathcal{Y}$ for the MLLM. Its output embeddings are passed to IMAGINER as textual conditions $\boldsymbol{c}_t$ for imagination, which are also used to generate verbal abduction results. On the right, IMAGINER is adapted from a text-to-image diffusion model through the integration of lightweight adapters. It takes $\boldsymbol{c}_t$ and $\boldsymbol{c}_v$ as multimodal conditions, where $\boldsymbol{c}_v$ is visual local-global hybrid representations extracted from the observations. The two components are trained end-to-end with $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{Diffusion}}$.

REASONER: Abduction in Verbal Mode

The REASONER enhances the abduction of MLLMs in verbal mode. It first generates candidate hypotheses using a blind LLM. Then, a causality-aware hypotheses selection module is proposed to prune visually irrelevant hypotheses based on causal relevance. The remaining hypotheses are subsequently passed to MLLMs as targeted priors for verbal abduction.

Causality-aware Hypotheses Generation (CHG)

Step 1: Candidate Verbal Hypotheses Generation. Abductive reasoning presents a significant challenge due to its vast and complex space of plausible explanations. To alleviate this, we leverage the knowledge-rich capabilities of advanced LLMs to generate a diverse set of candidate hypotheses, thereby narrowing the reasoning space. Specifically, we first employ a pre-trained MLLM to generate video captions for each observed video clip in $\mathcal{O}$, yielding a sequence of descriptions $\mathcal{C} = \{C_t\}_{t=1}^{T-1}$, with each C_t corresponds to the description of O_t. Afterwards, we prompt GPT-4o-mini (Hurst et al. 2024) to infer plausible missing events under the instruction: *You are an event-completion expert, infer the most plausible event at the [MASK] position.*. To ensure diversity and reduce redundancy, we query GPT-4o-mini multiple times per video instance using a relatively high sampling temperature (*e.g.*, 1.4). This finally results in a collection of L diverse candidate hypotheses denoted as $\mathcal{Y} = \{Y_i\}_{i=1}^{L}$, where each Y_i represents a distinct verbal explanation for the missing event.

Step 2: Causality-aware Cross-modal Hypotheses Selection. Since the hypotheses in $\mathcal{Y}$ are derived solely from textual captions, they may include low-quality, hallucinated candidates that hinder effective reasoning in MLLMs. Therefore, we introduce a contrastive learning based hypotheses selection module to identify causally relevant hypotheses from $\mathcal{Y}$ by leveraging $\mathcal{O}$. Unlike standard contrastive learning, which merely establishes superficial similarity between visual and textual modalities, this module is specifically designed to capture the causal relevance between visual observations and textual hypotheses.

Sufficient high-quality negative samples are crucial in contrastive learning (Chen et al. 2020; Robinson et al. 2021). While the ground-truth explanation E_h could directly serve as the positive hypothesis, constructing diverse and semantically meaningful negative hypotheses remains a challenge. To generate these negative hypotheses, we utilize GPT-4o-mini (Hurst et al. 2024). The prompt begins with a task description: *There is a contrastive learning task aimed at matching the missing video caption using a series of observed videos.* We then provide GPT with the positive explanation E_h, the observed captions $\mathcal{C}$, and instructions: *The negative samples should differ in semantics from the positive sample, but still align with the logical context of observed captions.* We call GPT-4o-mini multiple times to obtain M

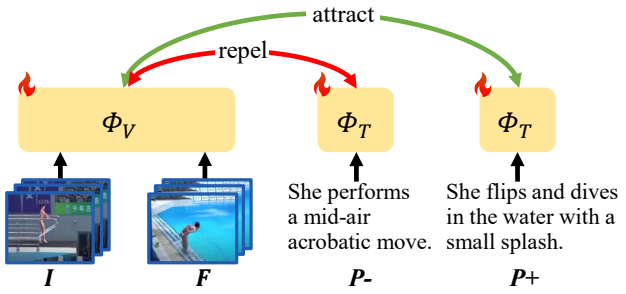

Figure 2: **Illustration of cross-modal causal contrastive learning.** Encoders Φ_V, Φ_T learn to attract causally plausible hypotheses $\mathcal{P}^+$ to the observations, and repel causally irrelevant hypotheses $\mathcal{P}^-$.

negative hypotheses.

Given any input $\mathcal{V}$, it can be naturally partitioned into three sequential segments: the *initial* segment $\mathcal{I}$, the *process* segment $\mathcal{P}$, and the *final* segment $\mathcal{F}$, and H may correspond to any of them. In the following, we take the case where H serves as the *process* segment as an illustrative example. The observations before and after H are then treated as the *initial* and *final* segments. As illustrated in Fig. 2, the module comprises a vision encoder Φ_V and a text encoder Φ_T, which project embeddings of these segments into a joint causal space. Specifically, Φ_V encodes the observed *initial* and *final* segments into visual embeddings $X_\mathcal{I}$ and $X_\mathcal{F}$. Φ_T encodes the postive and negative hypotheses into textual embedding $X_\mathcal{P}^+$ and $X_\mathcal{P}^-$.

During training, the model is optimized with a contrastive objective that maximizes the causal relevance between the observed video and the positive hypothesis, while minimizing the relevance with negative hypotheses. This is achieved via the NT-Xent loss (Chen et al. 2020):

$$\mathcal{L}_{\text{Contrast}} = -\log \frac{\exp\left(\langle \boldsymbol{X}_\mathcal{I} + \boldsymbol{X}_\mathcal{P}^+, \boldsymbol{X}_\mathcal{F} \rangle / \tau\right)}{\sum_{i=1}^{M} \exp\left(\langle \boldsymbol{X}_\mathcal{I} + \boldsymbol{X}_\mathcal{P}^{i-,+}, \boldsymbol{X}_\mathcal{F} \rangle / \tau\right)}, \quad (1)$$

where τ is the temperature coefficient, $X_\mathcal{P}^{i-}$ is the embedding of the i-th negative hypothesis, and $\langle \cdot, \cdot \rangle$ is cosine similarity between embeddings.

During inference, for each candidate hypothesis $Y_i \in \mathcal{Y}$ in Step 1, we project it into the joint space with Φ_T and compute its causal relevance score to the observed videos:

$$\text{Score}(Y_i) = \langle \boldsymbol{X}_\mathcal{I} + \boldsymbol{X}_{Y_i}, \boldsymbol{X}_\mathcal{F} \rangle, \quad (2)$$

where $\boldsymbol{X}_{Y_i} = \Phi_T(Y_i)$ denotes the embedding of Y_i. We then rank all candidate hypotheses based on their scores and select the top-k most causally aligned hypotheses for downstream reasoning.

Hypotheses Guided Verbal Abduction in MLLMs

To supply the MLLM with high-quality hypotheses, we retain only top-k highest-scoring hypotheses in the prompt for MLLM reasoning. We concatenate multiple events within $\mathcal{V}$ into a single video. For the unobserved event H, we fill the gap with placeholder frames with random pixels. To explicitly inform the MLLM of the position of H, we add numbers on video frames tailored for our VAR task. Instead of frame-level indices in Number-Prompt (Wu et al. 2025), these numbers encode event-level indices, allowing the model to understand the temporal ordering between different events. The output embeddings of REASONER (denoted by $\boldsymbol{c}_t$) is subsequently passed to IMAGINER as conditions for imagination. $\boldsymbol{c}_t$ is further decoded into verbal explanation as abduction result of **`AbductiveMLLM`**.

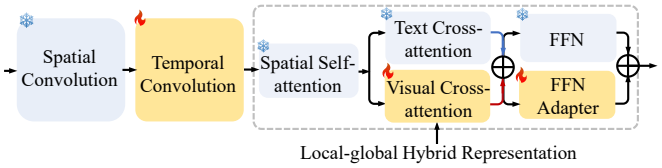

Figure 3: **A standard U-Net block of Stable Diffusion with proposed adapters.** During training, we only update parameters of the adapters (in yellow), and freeze parameters of other modules (in blue).

IMAGINER: Abduction in Pictorial Mode

The IMAGINER serves to provide richer contextual cues to facilitate the abduction of MLLMs in pictorial mode. Notably, as our objective centers on extracting visual guidance rather than optimizing video generation, we implement IMAGINER by extending a text-to-image diffusion model (*i.e.*, Stable Diffusion) to video generation through lightweight adapters. As shown in Fig. 3, we introduce three kinds of adapters to each U-Net block in Stable Diffusion.

Visual Cross-attention Adapter (V-Adapter). In vanilla Stable Diffusion, the U-Net attention blocks perform only self-attention for individual frames, neglecting the information from other frames. In our case, it fails to leverage the observed visual cues in $\mathcal{O}$, which potentially contain valuable information relevant to textual explanations. To address this issue, we introduce V-Adapter to inject informative visual priors into the model.

Directly incorporating all frames from $\mathcal{O}$, however, is computationally expensive and prone to noise due to redundancy. To overcome this, we propose an efficient strategy to extract a *local-global hybrid representation* from $\mathcal{O}$ that captures both fine-grained and holistic visual semantics relevant to the textual explanation E_h. *For the local representation*, we employ CLIP's image and text encoders to obtain embeddings $\{\boldsymbol{c}_v^i\}_{i=1}^N$ for all N frames in $\mathcal{O}$, and $\boldsymbol{c}_h$ for E_h. Then we calculate the similarity scores:

$$\gamma^i = \frac{\exp\left(\text{sim}(\boldsymbol{c}_v^i, \boldsymbol{c}_h)\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(\boldsymbol{c}_v^j, \boldsymbol{c}_h)\right)} \in [0, 1], \quad (3)$$

and only concatenate high-scoring frames together to form the local representation $\boldsymbol{c}_{local}$. *For the global representation*, we compute a weighted average of $\{\boldsymbol{c}_v^i\}_{i=1}^N$ based on the similarity scores $\{\gamma^i\}_{i=1}^N$, yielding the representation $\boldsymbol{c}_{global} = \sum_{i=1}^N \gamma^i \boldsymbol{c}_v^i$. Finally, we concatenate the local and global representations as the visual condition, denoted as $\boldsymbol{c}_v = [\boldsymbol{c}_{local}; \boldsymbol{c}_{global}]$. Hence, V-Adapter can be formulated

as:

$$\text{V-Adapter}(\boldsymbol{Q}, \boldsymbol{K}_v, \boldsymbol{V}_v) = \text{Softmax}\Big(\frac{\boldsymbol{Q}\,\boldsymbol{K}_v^\top}{\sqrt{d_k}}\Big)\,\boldsymbol{V}_v, \quad (4)$$

where $\boldsymbol{Q} = \boldsymbol{x}\boldsymbol{W}^q, \boldsymbol{K}_v = \boldsymbol{c}_v\boldsymbol{W}_v^k, \boldsymbol{V}_v = \boldsymbol{c}_v\boldsymbol{W}_v^v$. $\boldsymbol{K}_v$/$\boldsymbol{V}_v$ are the key/value in cross-attention computation of V-Adapter, $\boldsymbol{W}^q$/$\boldsymbol{W}_v^{k,v}$ are projection matrices.

We integrate the V-Adapter in parallel with the original text cross-attention in each U-Net block, while freezing the textual attention parameters. This enables the model to attend to both visual and textual cues simultaneously. The outputs from the parallel cross-attention branches are then summed. The process is depicted as follows:

$$\boldsymbol{x} = \text{CrossAttn}(\boldsymbol{Q}, \boldsymbol{K}_t, \boldsymbol{V}_t) + \text{V-Adapter}(\boldsymbol{Q}, \boldsymbol{K}_v, \boldsymbol{V}_v), \quad (5)$$

where $\boldsymbol{K}_t = \boldsymbol{c}_t\boldsymbol{W}_t^k$, $\boldsymbol{V}_t = \boldsymbol{c}_t\boldsymbol{W}_t^v$, $\boldsymbol{K}_t$/$\boldsymbol{V}_t$ are the key/value of the textual cross-attention, $\boldsymbol{W}_t^{k,v}$ are projection matrices.

Temporal Convolution Adapter (T-Adapter). The T-Adapter is designed to model temporal dependencies across frames and is appended after the spatial convolution layers in each U-Net block. It uses depth-wise 3D convolutional in a projected lower-dimensional space, which can alleviate the complexity of temporal modeling (Singer et al. 2022; Blattmann et al. 2023; Xing et al. 2024). To keep structural consistency and further improve temporal modeling, we adopt a fully convolutional design, which is defined as:

$$\text{T-Adapter}(\boldsymbol{x}) = \boldsymbol{x} + \text{Conv3D}_{up}\left(\text{Conv3D}_{down}(\boldsymbol{x})\right), \quad (6)$$

where Conv3D_{up} and Conv3D_{down} denote the up-projection and down-projection layers, both are 3D convolutions.

FFN Adapter (F-Adapter). The F-Adapter enhances spatial representation while preserving the integrity of the original feed-forward network. It is introduced as a parallel block to the FFN layer, making sure that the pretrained FFN remains unchanged while adapting to the spatial features of videos. FFN adapter consists of two fully connected (FC) layers with GELU activation (Xing et al. 2024), which can be formulated as:

$$\text{F-Adapter}(\boldsymbol{x}) = \boldsymbol{x} + \text{FC}_{up}(\text{GELU}(\text{FC}_{down}(\boldsymbol{x}))), \quad (7)$$

where FC_{up} and FC_{down} are the up-projection and down-projection layers.

Network Training

We adopt a two-stage training paradigm in which the modules are first trained independently before undergoing joint end-to-end optimization. In Stage I, for REASONER, the MLLM is finetuned with LoRA under the standard cross-entropy loss $\mathcal{L}_{CE}$, enabling it to generate plausible hypotheses from incomplete observations. For IMAGINER, we freeze the weights of stable diffusion and only update the parameters of adapters using the same conditional latent diffusion loss $\mathcal{L}_{Diffusion}$ in (Rombach et al. 2022). We also apply Min-SNR weighting strategy (Hang et al. 2023), which adaptively reweights the loss at each diffusion timestep to accelerate the convergence of the diffusion model. In Stage II, REASONER and IMAGINER are jointly tuned in an end-to-end manner. The overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{Diffusion}}, \quad (8)$$

where α is a coefficient that balances the two terms.

Experiment

Experimental Setup

Network Architecture. In REASONER, the MLLM is implemented as Qwen2VL-7B-Instruct (Wang et al. 2024a). Φ_V consists of pretrained ResNet200 (He et al. 2016)/BN-Inception (Ioffe and Szegedy 2015) as in (Liang et al. 2022) and trainable 2-layer Transformer encoder. Φ_T consists of pretrained CLIP text encoder and trainable 2-layer MLP. In IMAGINER, Stable Diffusion-v1-4 (Rombach et al. 2022) is the backbone, with 256×256 resolution and 32×32 latent size. The above two modules is connected by the Bridge Layer, which is implemented as a 2-layer MLP with SiLU activation.

Dataset. We validate our approach on two datasets:

- **VAR** (Liang et al. 2022). The VAR dataset contains 8,606 annotated samples sourced from 3,718 unique videos. Each video includes an average of 4.17 events, each lasting approximately 37.8 seconds. The dataset is split into train/val/test splits, containing 7,053/460/1,093 samples.
- **YouCookII** (Zhou, Xu, and Corso 2018). YouCookII is a large-scale cooking video dataset containing 1,333/457/210 videos for train/val/test. We follow the setup of (Tan et al. 2025b) to organize the videos for the VAR task. Specifically, we re-partition the original training and validation sets to obtain 1,533/257 videos for train/test. For each video, we iteratively select one event as the explanation event and treat the remaining events as observed events. Finally, we obtain 11,737/1,870 data samples for train/test.

Competitor. We present a comprehensive comparison between **AbductiveMLLM** and state-of-the-art models, including traditional specialized small-scale models (Liang et al. 2022; Tan et al. 2025b; Xu et al. 2024), proprietary and open-source MLLMs (Hurst et al. 2024; Li et al. 2024a; Wang et al. 2024a). Each traditional model is trained separately on `train` sets of VAR and YouCookII datasets. We also finetune Qwen2VL-7B-Instruct on the datasets for the same total epochs as our baseline (Qwen2VL-7B$^{\text{FT}}$).

Training Configuration. In Stage I, we first finetune the MLLM and the diffusion model separately for 2 epochs. To train the cross-modal constrastive learning module, we generate 100 hard negatives for each positive E_h. Then we train it for 10 epochs and use the model with the highest training accuracy to prune candidate hypotheses. We set $k = 3$ in the top-k hypotheses selection. In Stage II, we finetune the REASONER and IMAGINER in an end-to-end manner for 1 epoch. All models are trained on 4 A800 GPUs with 80GB memory per-card.

Metric. We follow prior works (Liang et al. 2022; Tan et al. 2025b; Xu et al. 2024) to use five well-known automated metrics *i.e.*, BLEU@4, METEOR, ROUGE-L, CIDEr and BERT-S (Zhang et al. 2019) for evaluation.

Main Result

Quantitative Result. As shown in Table 1 and Table 2, **AbductiveMLLM** achieves the best performance across all

Method	BLEU@4	METEOR	ROUGE	CIDEr	BERT-S
Human	11.35	19.36	36.92	147.79	40.59
• **Traditional Models**					
VTrans (Zhou et al. 2018)	0.71	6.92	19.12	7.11	22.13
MFT (Xiong, Dai, and Lin 2018)	1.81	7.16	19.16	17.67	25.90
Trans-XL (Dai et al. 2019)	2.96	7.51	20.94	24.54	27.23
MART (Lei et al. 2020)	2.86	7.47	20.87	24.05	27.77
PDVC (Wang et al. 2021)	3.00	8.54	20.71	25.14	27.80
REASONER (Liang et al. 2022)	3.44	9.05	22.89	30.75	30.64
KN-VLM (Tan et al. 2025b)	4.72	10.74	24.40	37.20	33.17
UPD-Trans (Xu et al. 2024)	5.40	11.16	25.62	41.66	30.80
• **MLLMs**					
GPT-4o-mini (Hurst et al. 2024)	0.63	7.38	13.64	7.30	12.27
VideoChat2-7B (Li et al. 2024a)	1.24	7.55	17.06	19.51	26.40
Qwen2VL-7B (Wang et al. 2024a)	2.41	11.29	21.61	29.25	30.01
Qwen2VL-7B$^{\texttt{FT}}$ (Wang et al. 2024a)	5.67	12.77	27.11	50.82	36.03
`AbductiveMLLM`	**6.54**	**13.41**	**27.95**	**57.04**	**36.80**

Table 1: Quantitative results on VAR `test`. `FT` means the model is finetuned on the dataset.

Method	BLEU@4	METEOR	ROUGE	CIDEr	BERT-S
• **Traditional Models**					
REASONER (Liang et al. 2022)	3.54	9.47	24.62	32.99	23.19
• **MLLMs**					
GPT-4o-mini (Hurst et al. 2024)	0.45	4.22	12.78	14.15	9.75
VideoChat2-7B (Li et al. 2024a)	0.49	4.59	14.31	17.82	7.96
Qwen2VL-7B (Wang et al. 2024a)	2.46	8.41	22.10	35.83	21.83
Qwen2VL-7B$^{\texttt{FT}}$ (Wang et al. 2024a)	5.66	12.62	28.64	68.44	29.09
`AbductiveMLLM`	**6.16**	**13.46**	**30.06**	**77.70**	**30.77**

Table 2: Quantitative results on YouCookII `test`. `FT` means the model is finetuned on the dataset.

metrics on both VAR `test` and YouCookII `test` benchmarks, demonstrating consistent advantages over both traditional models and advanced MLLMs.

Several key observations can be drawn from the results. First, compared to the best traditional model (*i.e.*, UPD-Trans), our method achieves significantly higher scores across all metrics on VAR, including **+1.14** BLEU@4, **+2.25** METEOR, **+2.33** ROUGE, **+15.38** CIDEr, **+6.00** BERT-S. This demonstrates that MLLMs, when equipped with abductive reasoning module, can outperform task-specific architectures, offering a more scalable solution for VAR. Second, our model also surpasses all zero-shot MLLMs (*i.e.*, GPT-4o-mini, VideoChat2-7B, Qwen2VL-7B), with improvements of over **+6.34** ROUGE, **+27.79** CIDEr, **+6.79** BERT-S on VAR, and **+7.96** ROUGE, **+41.87** CIDEr, **+8.94** BERT-S on YouCookII. These results reveals the lack of abductive reasoning capabilities in existing general LLMs. Even when compared to an MLLM specifically fine-tuned for the VAR task (Qwen2VL-7B$^{\texttt{FT}}$), our method achieves consistent gains on both datasets (*e.g.*, **+0.84** ROUGE, **+6.22** CIDEr on VAR, and **+1.42** ROUGE, **+9.26** CIDEr on YouCookII), confirming that the integration of verbal and pictorial abduction provides complementary VAR capability beyond tuning alone can offer. Third, we observe that there still remains a significant gap between human abduction and AI models, which indicates that AI models still have substantial room to grow before reaching human-level cognitive capabilities.

Qualitative Result. Fig. 4 contains the explanatory hypotheses from **`AbductiveMLLM`** and other competitors (Liang et al. 2022; Hurst et al. 2024; Wang et al. 2024a) as well as groundtruth sentences. We can find that our method is able to discover and correctly describe the cause-effect chain, and hence generate a plausible hypothesis: '`throws the frisbee again and the dog catches it`', that well explans the observed events. In contrast, other competitors typically produce unsatisfactory results, *e.g.*, REASONER (Liang et al. 2022) misidentifies the person's gender, GPT-4o-mini offers only *general descriptions* rather than *fine-grained reasoning*, baseline (Qwen2VL-7B$^{\texttt{FT}}$) fails to infer the man's action.

Ablation Study

We conduct ablative experiments on VAR `test` for in-depth analyzing each design in our approach.

Key Component Analysis. We first study the efficacy of core model designs in Table 3. The first row gives the performance of the baseline (Qwen2VL-7B$^{\texttt{FT}}$). The results in the second and third rows reveal that both CHG and IMAGINER can improve all five metrics, and the larger gains are achieved on overlap-sensitive scores (BLEU@4,

Groundtruth: [The man throws the frisbee, the dog brings it back.] [The man throws it again and the dog returns it.] [The man wipes the frisbee off.]
REASONER: She throws a frisbee, and the dog catches it.
GPT-4o-mini: The dog and its owner were engaged in training or playing with a frisbee, interspersed with moments of interaction and commands between the owner and the dog.
Baseline (Qwen2VL−7B^{FT}): The dog runs after the frisbee and brings it back to the trainer.
AbductiveMLLM (Ours): The man throws the frisbee again and the dog catches it.

Figure 4: Qualitative comparison of **AbductiveMLLM** on an example from VAR `test`.

CHG	IMAGINER	BLEU@4	METEOR	ROUGE	CIDEr	BERT-S
		5.67	12.77	27.11	50.82	36.03
✓		6.33	12.96	27.21	53.60	36.31
	✓	6.35	13.07	27.52	55.00	36.40
✓	✓	**6.54**	**13.41**	**27.95**	**57.04**	**36.80**

Table 3: Diagnostic experiments for **AbductiveMLLM**.

k	BLEU@4	METEOR	ROUGE	CIDEr	BERT-S
0	6.35	13.07	27.52	55.00	36.40
1	6.41	12.94	27.52	54.15	36.31
3	**6.54**	**13.41**	**27.95**	**57.04**	**36.80**
6	6.32	13.13	27.68	54.89	36.47
10	6.29	13.22	27.66	53.66	36.40

Table 4: Ablation study on top-k hypotheses selection.

α (Eq. 8)	BLEU@4	METEOR	ROUGE	CIDEr	BERT-S
1	6.33	13.36	27.86	54.91	36.76
3	6.52	13.39	**28.11**	56.54	**36.86**
5	**6.54**	**13.41**	27.95	**57.04**	36.80
7	6.42	13.40	27.90	55.32	36.75
9	6.50	13.32	27.86	55.52	36.65

Table 5: Ablation study on α.

Variant	BLEU@4	METEOR	ROUGE	CIDEr	BERT-S
AbductiveMLLM	**6.54**	**13.41**	**27.95**	**57.04**	**36.80**
w/o V-Adapter	6.36	13.29	27.76	54.51	36.68
w/o T-Adapter	6.42	13.28	27.74	54.99	36.68
w/o F-Adapter	6.47	13.31	27.70	54.52	36.63

Table 6: Ablation study on IMAGINER adapters.

CIDEr), indicating that both designs can help the model produce essential content words. Furthermore, by comparing the second and third rows, IMAGINER can bring more gains on semantics-oriented metrics (METEOR, ROUGE) and the embedding-based BERT-S, suggesting richer, visually grounded verbal results. From the last row, we can conclude that combining the two designs together leads to the best results.

Top-k Hypotheses Selection. As we select the top-k candidate hypotheses from GPT, we study the impact of k in the hypotheses selection. As shown in Table 4, $k = 0$ means training without any hypothesis from GPT. We observe that when k is larger than 3, excessive candidate hypotheses leads to a noticeable drop in performance across all metrics; when k is less than 3, insufficient hypotheses fails to provide noticeable gains in performance. We therefore use $k = 3$ for all experiments.

Coefficient α. We study the impact of α in Eq. 8 in Table 5. A larger α indicates a greater degree of intervention by imagination in the training. Among the various α values we examined, the best performance is reached at $\alpha = 5$. Nonetheless, **AbductiveMLLM** maintains relatively stable performance with different α values, indicating that our model's performance is not sensitive to the selection of α.

Adapters in IMAGINER. We assess the contribution of the proposed adapters in IMAGINER. As shown in Table 6, individually removing each adapter from IMAGINER leads to a performance drop, especially on the CIDEr metric, which indicates the model produces fewer accurate key terms in verbal abduction. The results confirm the necessity of each adapter in IMAGINER. Moreover, training with variants of IMAGINER still outperforms the model without IMAGINER (the second row in Table 3), which further proofs the importance of pictorial thinking.

Conclusion

In this work, we present **AbductiveMLLM**, the pioneer to enhance the abductive reasoning capabilities of MLLMs from verbal and pictorial perspectives. We propose a causality-aware contrastive learning algorithm to mine hypotheses with high causal relevance, reducing reasoning space and providing verbal priors for MLLMs (REASONER). Unlike prior methods that rely solely on verbal abduction, we incorporate pictorial thinking via an adapted diffusion model (IMAGINER). By jointly training the two components, our method effectively emulates the human-like interplay between language and imagination. Extensive experiments on standard benchmarks show that our method consistently outperforms existing small-scale models and competitive MLLM baselines, setting the new state-of-the-art.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62576035), Beijing Natural Science Foundation (L252036), and CAAI-Lenovo Blue Sky Research Fund.

References

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*.

Chen, F.; Han, M.; Zhao, H.; Zhang, Q.; Shi, J.; Xu, S.; and Xu, B. 2023. X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages. *arXiv preprint arXiv:2305.04160*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; and Bing, L. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Chinchure, A.; Ravi, S.; Ng, R.; Shwartz, V.; Li, B.; and Sigal, L. 2024. Black Swan: Abductive and Defeasible Video Reasoning in Unpredictable Events. *arXiv preprint arXiv:2412.05725*.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*.

Hang, T.; Gu, S.; Li, C.; Bao, J.; Chen, D.; Hu, H.; Geng, X.; and Guo, B. 2023. Efficient diffusion training via min-snr weighting strategy. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hessel, J.; Hwang, J. D.; Park, J. S.; Zellers, R.; Bhagavatula, C.; Rohrbach, A.; Saenko, K.; and Choi, Y. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *ECCV*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Lei, J.; Wang, L.; Shen, Y.; Yu, D.; Berg, T. L.; and Bansal, M. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*.

Li, M.; Han, K.; Xu, J.; Li, Y.; Wu, T.; Zhao, Z.; Miao, J.; Zhang, S.; and Chen, J. 2024b. Cross-modal Observation Hypothesis Inference. In *ACM MM*.

Li, M.; Wang, T.; Xu, J.; Han, K.; Zhang, S.; Zhao, Z.; Miao, J.; Zhang, W.; Pu, S.; and Wu, F. 2023. Multi-modal action chain abductive reasoning. In *ACL*.

Liang, C.; Wang, W.; Zhou, T.; and Yang, Y. 2022. Visual abductive reasoning. In *CVPR*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Empirical Methods in Natural Language Processing*.

Luo, R.; Zhao, Z.; Yang, M.; Dong, J.; Qiu, M.-H.; Lu, P.; Wang, T.; and Wei, Z. 2023. Valley: Video Assistant with Large Language model Enhanced abilitY. *arXiv preprint arXiv:2306.07207*.

Lyu, C.; Wu, M.; Wang, L.; Huang, X.; Liu, B.; Du, Z.; Shi, S.; and Tu, Z. 2023. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. *arXiv preprint arXiv:2306.09093*.

Ma, S.; Ge, Y.; Wang, T.; Guo, Y.; Ge, Y.; and Shan, Y. 2025. GenHancer: Imperfect Generative Models are Secretly Strong Vision-Centric Enhancers. *CoRR*.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *ACL*.

Peirce, C. S. 1931. *Collected papers of charles sanders peirce*. Harvard University Press.

Robinson, J.; Chuang, C.-Y.; Sra, S.; and Jegelka, S. 2021. CONTRASTIVE LEARNING WITH HARD NEGATIVE SAMPLES. In *ICLR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Shanahan, M. 2005. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science*, 29(1): 103–134.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *ICLR*.

Tan, C.; Yeo, C. K.; Tan, C.; and Fernando, B. 2025a. Inferring Past Human Actions in Homes with Abductive Reasoning. In *WACV*.

Tan, K.; Qi, Z.; Zhong, J.; Xu, Y.; and Zhang, W. 2025b. KN-VLM: KNowledge-guided Vision-and-Language Model for visual abductive reasoning. *Multimedia Systems*, 31(2): 146.

Thagard, P.; and Shelley, C. 1997. Abductive reasoning: Logic, visual thinking, and coherence. In *CLMPST*, 413–427.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Q.; Yu, Y.; Yuan, Y.; Mao, R.; and Zhou, T. 2025a. VideoRFT: Incentivizing Video Reasoning Capability in MLLMs via Reinforced Fine-Tuning. *arXiv preprint arXiv:2505.12434*.

Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; and Luo, P. 2021. End-to-end dense video captioning with parallel decoding. In *ICCV*.

Wang, W.; Sun, Q.; Zhang, F.; Tang, Y.; Liu, J.; and Wang, X. 2025b. Diffusion feedback helps clip see better. *ICLR*.

Wang, Y.; Chen, W.; Han, X.; Lin, X.; Zhao, H.; Liu, Y.; Zhai, B.; Yuan, J.; You, Q.; and Yang, H. 2024b. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Wang, C.; Chen, G.; Pei, B.; Zheng, R.; Xu, J.; Wang, Z.; et al. 2024c. Internvideo2: Scaling video foundation models for multimodal video understanding. In *ECCV*.

Wang, Y.; Zeng, Y.; Zheng, J.; Xing, X.; Xu, J.; and Xu, X. 2024d. VideoCoT: A Video Chain-of-Thought Dataset with Active Annotation Tool. In *ACL Workshop*.

Wu, J.; Gan, W.; Chen, Z.; Wan, S.; and Yu, P. S. 2023. Multimodal large language models: A survey. In *IEEE BigData*.

Wu, Y.; Hu, X.; Sun, Y.; Zhou, Y.; Zhu, W.; Rao, F.; Schiele, B.; and Yang, X. 2025. Number it: Temporal Grounding Videos like Flipping Manga. In *CVPR*.

Xing, Z.; Dai, Q.; Hu, H.; Wu, Z.; and Jiang, Y.-G. 2024. Simda: Simple diffusion adapter for efficient video generation. In *CVPR*.

Xiong, Y.; Dai, B.; and Lin, D. 2018. Move forward and tell: A progressive generator of video descriptions. In *ECCV*.

Xu, H.; Ye, Q.; Wu, X.-W.; Yan, M.; Miao, Y.; Ye, J.; Xu, G.; Hu, A.; Shi, Y.; Xu, G.; Li, C.; Qian, Q.; Que, M.; Zhang, J.; Zeng, X.; and Huang, F. 2023a. Youku-mPLUG: A 10 Million Large-scale Chinese Video-Language Dataset for Pre-training and Benchmarks. *arXiv preprint arXiv:2306.04362*.

Xu, M.; Jiang, G.; Liang, W.; Zhang, C.; and Zhu, Y. 2023b. Active reasoning in an open-world environment. *NeurIPS*.

Xu, W.; Miao, Z.; Tian, Y.; Cen, Y.; Wan, L.; and Xiaole, M. 2024. Probabilistic Distillation Transformer: Modelling Uncertainties for Visual Abductive Reasoning. In *ACM MM*.

Zhang, H.; Ee, Y. K.; and Fernando, B. 2024. RCA: Region Conditioned Adaptation for Visual Abductive Reasoning. In *ACM MM*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, W.; Rao, Y.; Tang, Y.; Zhou, J.; and Lu, J. 2022. Videoabc: A real-world video dataset for abductive visual reasoning. *IEEE Trans. Image Process.*, 31: 6048–6061.

Zhou, L.; Xu, C.; and Corso, J. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI*.

Zhou, L.; Zhou, Y.; Corso, J. J.; Socher, R.; and Xiong, C. 2018. End-to-end dense video captioning with masked transformer. In *CVPR*.

Zohar, O.; Wang, X.; Dubois, Y.; Mehta, N.; Xiao, T.; Hansen-Estruch, P.; Yu, L.; Wang, X.; Juefei-Xu, F.; Zhang, N.; Yeung-Levy, S.; and Xia, X. 2024. Apollo: An Exploration of Video Understanding in Large Multimodal Models. *arXiv preprint arXiv:2412.10360*.