MIMIC-BENCH: EXPLORING THE USER-LIKE THINKING AND MIMICKING CAPABILITIES OF MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid advancement of multimodal large language models (MLLMs) has greatly prompted the video interpretation task, and numerous works have been proposed to explore and benchmark the cognition and basic visual reasoning capabilities of MLLMs. However, practical applications on social media platforms demand MLLMs that can emulate user-like thinking and behavior when interpreting user-generated videos, which has been rarely studied in current research. To bridge the gap and get closer to general practical artificial intelligence (AI), we first construct **MIMIC-Data**, a large-scale dataset containing 150K+ user-shared videos with corresponding information including captions, tags, comments, *etc.* Then, we present **MIMIC-Bench**, a large-scale benchmark building upon curated 4,000 user-shared videos from MIMIC-Data, which is designed to evaluate user-like thinking and mimicking capabilities of MLLMs in real-world video contexts. MIMIC-Bench not only supports user-like thinking challenges including creator intent, user feedback interpretation, *etc.*, but also introduces a novel comment imitation task to assess whether MLLMs can generate human-like responses to video content. Based on MIMIC-Data and MIMIC-Bench, we develop **MIMIC-Chat**, which integrates spatial and temporal features into a large language model, and finetunes the model to perform user-like thinking and mimicking tasks. Extensive experiments conducted based on 24 existing MLLMs and our MIMIC-Chat model that current MLLMs exhibit limited capabilities to perform human-like thinking and responses, and MIMIC-Chat performs better to some extent. We hope MIMIC-Bench can contribute to the advancement of human-aligned video understanding in the multi-modal era. The MIMIC-Data, MIMIC-Bench, and MIMIC-Chat will be released upon the publication.

1 INTRODUCTION

In recent years, video platforms have rapidly emerged as mainstream media for social interaction and knowledge sharing. Beyond passive viewing, users actively participate in shaping the video ecosystem through content creation, commenting, and engagement. A single user-generated video is often accompanied by rich metadata and viewer interactions, such as titles, descriptions, tags, and comments, *etc.*, which reflect not only the video content but also how humans perceive, interpret, and respond to it. These contents embody high-level human cognition and social behavior, offering a rich yet underexplored resource for evaluating machine intelligence. Studying how multi-modal large language models (MLLMs) interpret such human-centric signals is crucial for advancing toward general-purpose AI systems that can truly understand and engage into social and cultural contexts.

Driven by the effectiveness of integrating multimodal information into large language models (LLMs) to acquire the scene perception ability (Radford et al., 2021; Li et al., 2021; Chen et al., 2020; Wang et al., 2022; Cho et al., 2021; Wang et al., 2021), recent progress in MLLMs (Zhang et al., 2025; Bai et al., 2025; Gao et al., 2024; Chen et al., 2024a) have led to strong performance on tasks involving visual understanding, temporal reasoning, and open-ended language generation. Powered by scalable visual encoders and multi-frame alignment mechanisms, these models achieve impressive results on several benchmarks (Li et al., 2024b; Fu et al., 2024; Mangalam et al., 2023), *etc.*, which assess the capabilities of MLLMs in video captioning, action recognition, and event prediction. However, these benchmarks typically rely on curated visual-only datasets and designer-crafted questions, focusing on low- or mid-level perception tasks that emphasize what happens in the video while neglecting how humans think , feel, and react. The lack of real-world user context and organically generated content limits their capacity to evaluate whether MLLMs can simulate human

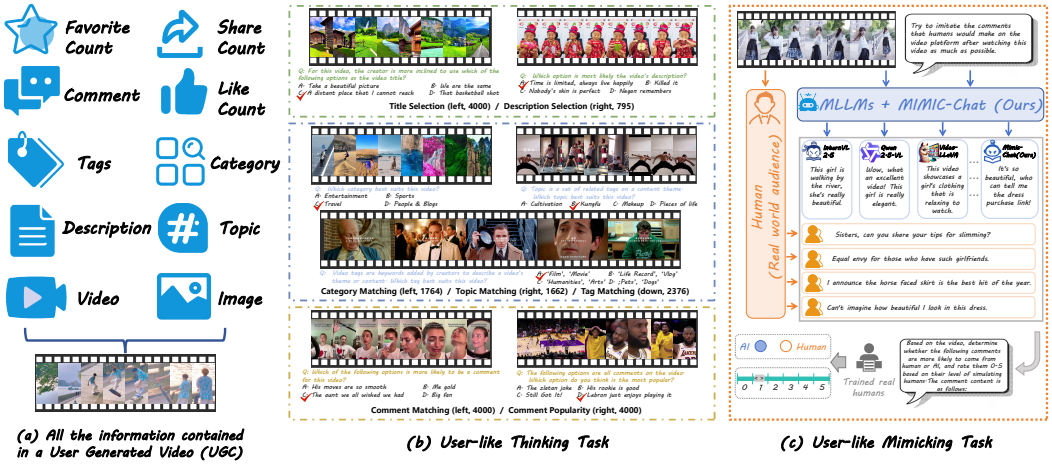

Figure 1: Overview of **MIMIC-Bench**. **(a)** Left: Multi-source metadata from user-shared videos. **(b)** Center: **User-like thinking task**: across three axes—CIU, CAM, and UIU. **(c)** Right: **User-like mimicking task**: human-like comment simulation pipeline.

cognition and language behavior in authentic video scenarios. Although some studies have begun to explore emotions in multimodal data such as EmoLLM (Yang et al., 2024), *etc.*, the practical applications are more complex than simple emotional tasks.

To bridge this gap, we first introduce **MIMIC-Data**, which comprises 150K+ user-shared videos and corresponding metadata information, as shown in Figure 1(a). Then, we construct **MIMIC-Bench**, a large-scale benchmark designed to evaluate user-like thinking and behavioral capabilities of MLLMs in real-world, user-centric video applications. Unlike prior benchmarks that focus primarily on visual recognition and rely on synthetic or designer-curated questions, MIMIC-Bench is grounded in real user-generated content and targets the human-like video understanding, *i.e.*, how creators and viewers interpret, react to, and communicate about video content. As shown in Figure 1(b)(c), the benchmark consists of two components: (1) a User-like Thinking Task that spans three structured reasoning axes, including Creator Intent Understanding (CIU), Content Attribute Matching (CAM), and **User Interaction Understanding (UIU)**, and (2) a **User-like Mimicking Task** that evaluates whether MLLMs can generate and identify human-like comments on video content. Notably, MIMIC-Bench shifts the focus from factual QA to human-centered cognitive tasks, especially comment imitation, which directly evaluate whether MLLMs can approximate human-like reasoning and expression.

We also introduce **MIMIC-Chat**, a multi-modal model designed to simulate human thinking and expression. Built upon the spatial and temporal visual encoders and a large language model backbone, it is trained on over 150k video samples from the training set of MIMIC-Data, enabling joint learning of video semantics and human-style responses. In summary, our key contributions are:

- We construct **MIMIC-Data**, a large-scale dataset containing all metadata information for over 150K user-shared videos.
- We propose **MIMIC-Bench**, the first large-scale benchmark designed to evaluate human-aligned reasoning and communicative behavior of MLLMs in practical video understanding. Especially, we introduce a novel **comment imitation task** that offers a new lens to assess human-like thinking and mimicking capabilities of MLLMs.
- We develop **MIMIC-Chat**, a multi-modal model trained on 150k+ real-world samples, enabling the joint modeling of video semantics and social cognition.
- We conduct extensive experiments on MIMIC-Bench, and compare the performance of 24 state-of-the-art MLLMs and our MIMIC-Chat. Results show that current MLLMs exhibit limited capabilities on MIMIC-Bench, and MIMIC-Chat performs better to some extent.

2 Related Work

2.1 Multi-modal Large Language Models for Video Understanding

Recent advances in large language models (LLMs) (Eysenbach et al., 2023; Chiang et al., 2023) have shown strong performance in open-ended reasoning, contextual understanding, and instruction

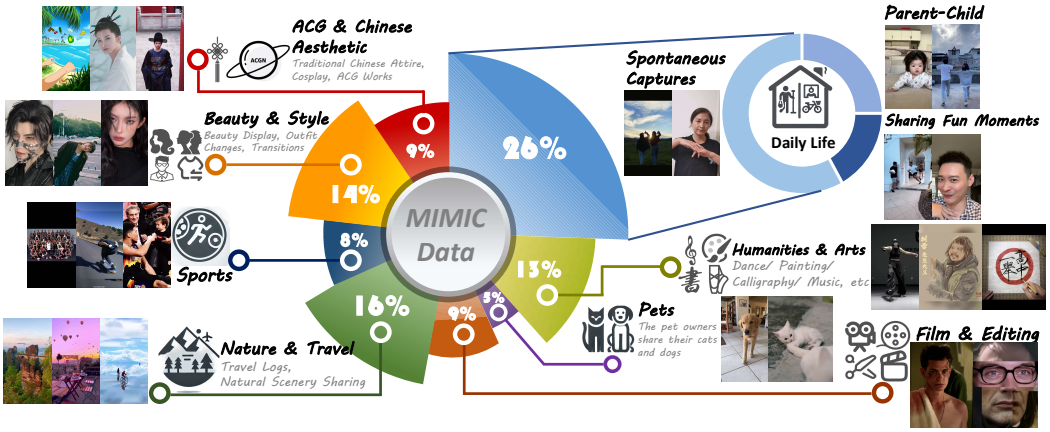

Figure 2: **Mimic-Data Overview**. We select 150,000+ videos, and each video is guaranteed to be of high quality and in line with public aesthetics. The videos can be categorized into 8 categories (Daily Life can be divided into: Spontaneous Captures, Parent-Child, and Sharing Fun Moments).

following (Radford et al., 2019; Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023). Building on this progress, multi-modal large language models (MLLMs) (Dai et al., 2023; OpenAI, 2023; Team et al., 2023; Zhu et al., 2023) have emerged to jointly process visual and textual inputs. Early works (Alayrac et al., 2022; Chen et al., 2022) introduced vision-language pretraining, while recent models (Liu et al., 2023; **?**; Sun et al., 2024) improved image-language alignment. As video becomes more prominent, MLLMs have been extended to handle temporal dynamics and multi-frame reasoning. Representative models like Video-LLaMA (Zhang et al., 2023; Cheng et al., 2024; Zhang et al., 2025), InternVL (Chen et al., 2023; 2024b; Gao et al., 2024; Chen et al., 2024a), and the Qwen-VL series (Bai et al., 2023; Wang et al., 2024a; Bai et al., 2025) combine visual encoders with language backbones to perform well on video captioning, action localization, and Q&A, leveraging spatiotemporal pooling, contrastive learning, and dense alignment. Despite excelling at factual understanding, most MLLMs fall short in modeling human-like cognition. These gaps highlight the need reflecting real-world user intent and communication and motivates our work.

2.2 Video Benchmarks and Evaluation Protocols

Numerous benchmarks (Fu et al., 2023; Li et al., 2023; Yu et al., 2023; Liu et al., 2025; Ning et al., 2023; Wang et al., 2024b) and datasets (Marino et al., 2019; Goyal et al., 2017; Song et al., 2024) have been developed to evaluate MLLMs on video understanding. Early works (Jang et al., 2017; Chowdhury et al., 2018) focus on short-form video QA, testing recognition of visual entities and actions. Recent benchmarks (Li et al., 2024b; Fu et al., 2024; Mangalam et al., 2023; Wu et al., 2024) extend evaluation to multimodal alignment, instruction following, long-range modeling, and multi-turn dialogue. These resources have improved content-level evaluation, primarily assessing "what happens in the video" via structured queries, while overlooking how humans interpret, react to, or linguistically engage with videos. Most evaluations stress visual recognition but neglect emotional and linguistic engagement with videos, while EmoLLM (Yang et al., 2024) touches affective reasoning yet leaves human-style commenting, social interaction, and expressive realism underexplored. User-generated signals like comments, descriptions, and feedback are rarely included, despite their relevance on real-world platforms. This limits assessment of models' ability to simulate human cognition and communication in natural contexts. In contrast, our benchmark incorporates real user content and authentic responses to evaluate MLLMs from a human-aligned perspective.

3 Benchmark Construction – MIMIC-Bench

In this section, we introduce **MIMIC-Bench**, a benchmark for evaluating human-aligned reasoning and communicative behavior in real-world, user-centric video scenarios. It follows three core principles: **Realism**—deriving tasks from authentic human interactions rather than synthetic prompts; **Task Orientation**—designing evaluations around multiple-choice and generation tasks with practical grounding; and **Human Alignment**—measuring models' ability to simulate human cognition and expression in the video domain.

3.1 Dataset Overview

Data Source and Initial Collection. To construct MIMIC-Bench, we first collect over 150,000 user-generated videos from mainstream short video platforms (TikTok and YouTube), which form

the foundation of **MIMIC-Data**. This dataset spans diverse topics such as lifestyle, education, travel, beauty, and humor (Figure 2). Each video is paired with metadata, including titles, tags, descriptions, comments, and interaction statistics, all sourced from publicly accessible content in compliance with platform policies. To ensure quality and usability, we applied initial filtering to remove videos with low resolution, excessive noise, or corruption, as well as those lacking meaningful metadata (*e.g.*, empty titles, no comments, very short duration). We prioritized samples with rich viewer engagement—measured by comment volume and like count—ensuring that selected videos reflect natural and diverse human responses. This high-quality pool forms the basis for both benchmark construction and model training.

Multi-Source Metadata Composition. Each video is represented as a multi-modal, multi-field unit combining content with surrounding human interactions. Beyond the visual and audio signals, each sample includes user-level metadata: title (creator intent), tags and topic labels (categorical annotations), description (context or emotional framing), category, user comments (subjective impressions), and popularity indicators (like counts). These components capture not only what the video presents but also how creators frame it and how viewers respond cognitively and emotionally. Unlike datasets focused solely on video-text pairs, our structure reflects the full communication loop of online video ecosystems.

Benchmark Subset Construction. To build tasks, we further filtered the 150,000 videos using a ranking score based on engagement metrics (likes, favorites, shares), publisher influence, and thematic relevance. We selected the **top 2%** from TikTok and **top 5%** from YouTube, yielding 4,000 highly engaging and representative samples.

These videos combine high content quality with rich user interaction and serve as the core resource for constructing challenging, cognitively demanding evaluation tasks.

To ensure reliability, we implemented a three-step quality control process: selecting highly interactive videos with rich viewer feedback, carefully designing distractors to avoid ambiguity, and conducting a large-scale manual review of over 20,000 QA entries from 4,000 videos. This procedure guarantees accuracy, coherence, and unambiguous semantics across the benchmark. The distribution of tasks reflects real-world user behavior rather than being artificially balanced, ensuring that the benchmark maintains ecological validity and mirrors authentic application scenarios.

Fairness and Validity To enhance fairness and temporal validity, we further filtered out samples that relied on ephemeral trends, niche memes, or culture-specific references. By retaining only general and widely comprehensible semantics—such as emotional tone, social context, and stylistic cues—MIMIC-Bench remains accessible across cultural backgrounds and stable over time.

3.2 Task Suite Design

3.2.1 User-like Thinking Task

To assess whether MLLMs can reason about video content in a human-aligned way, we design seven single-choice tasks grouped into three axes: Creator Intent Understanding (CIU), Content Attribute Matching (CAM), and User Interaction Understanding (UIU). These tasks leverage rich metadata from **MIMIC-Data** to evaluate both perceptual and cognitive understanding. These tasks are grounded in rich video metadata and evaluate both perceptual and cognitive understanding. Detailed information about the MIMIC-Bench benchmark, are provided in Appendix.

Creator Intent Understanding (CIU). This axis includes two sub-tasks. In *Title Selection*, the model selects the title most likely assigned by the creator, based on thematic and stylistic cues. In *Description Selection*, it identifies the description best aligned with the video's content and intent. Ground-truths come from original metadata; distractors are sampled from unrelated videos to ensure semantic contrast and minimize stylistic leakage.

Content Attribute Matching (CAM). This axis comprises three tasks. In *Tag Matching*, the model selects the tag that best captures content-level semantics. *Topic Matching* requires identifying the most relevant thematic label, while *Category Matching* asks for the correct predefined content category. All answers are derived from original annotations, and distractors are drawn from distinct videos with non-overlapping labels.

User Interaction Understanding (UIU). This axis focuses on modeling viewer behavior. In *Comment Matching*, the model selects the comment most likely reflecting genuine viewer feedback; the ground-truth is the top-liked comment, and distractors come from unrelated categories. In *Comment Popularity*, the model chooses the most-liked comment among four from the same video (top-1, top-10, top-50, top-100), based on subtle linguistic and contextual signals. All tasks follow a four-way

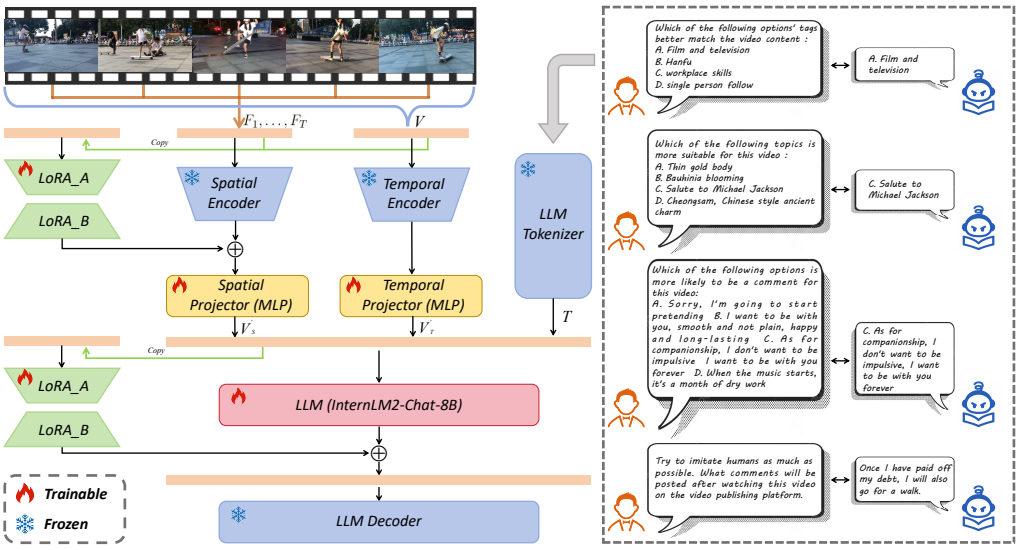

Figure 3: Overview of the proposed **MIMIC-Chat** architecture. Given sampled video frames, spatial and temporal features are extracted using dual encoders and projected into language space via lightweight MLPs. The projected visual tokens and task-specific instructions are fused and processed by a LoRA-enhanced InternLM2-Chat-8B model. The unified interface supports both structured reasoning tasks (*e.g.*, classification) and open-ended comment generation, enabling human-aligned video understanding.

single-choice format with one correct answer. The design emphasizes semantic confusion, distractor diversity, and option balance, ensuring that task success reflects genuine understanding rather than superficial cues or memorization. While some metadata-based tasks might appear straightforward, the overall suite is cognitively demanding. In particular, comment-related tasks (e.g., Comment Matching and Comment Popularity) require nuanced reasoning about social interaction, linguistic appeal, and collective preferences, making them non-trivial and discriminative.

3.2.2 User-like Mimicking Task

Beyond structured reasoning, we introduce a novel task—**Comment Imitation**—to evaluate whether MLLMs can simulat human-like comments in social video contexts. This task emphasizes linguistic creativity, emotional expression, and social reasoning rather than factual correctness.

Task Motivation. As the second core component of MIMIC-Bench, this task captures the imaginative, affective, and context-aware nature of real viewer comments. It evaluates capability: generating authentic, human-style comments. Although the task inevitably involves subjective judgment, we explicitly design the protocol and aggregation strategy to ensure reproducibility and comparability, transforming subjective perception into a systematic evaluation dimension.

Task Setup. For each of the 4,000 benchmark videos, we collect the **top-5** most-liked real comments. Each of the **24 MLLMs** then generates one comment per video, producing a pool of 5 real and 24 model-generated comments, anonymized and randomly shuffled.

Human Evaluation. We conduct al human study using the same setup to provide a perceptual reference. These annotations serve as both a performance upper bound and an indicator of which AI-generated comments "deceive" human judges.

Evaluation Metrics. We use two metrics: **Imitation Quality**, based on the percentage of comments judged as human and their average realism scores (0–5). Beyond serving as a benchmark component, this protocol and its associated metrics also provide a methodological contribution, offering a reusable framework for future studies on human-likeness evaluation in multimodal generation.

This task creates a unique loop—*generation* → *judgment* → *scoring*—offering a lens to assess models' social and linguistic alignment with humans in multimodal settings.

4 MIMIC-Chat

In this section, we present **MIMIC-Chat**, a unified multimodal framework tailored for human-aligned video understanding. The model supports both **User-like Thinking Tasks** and **User-like**

Mimicking Tasks in **MIMIC-Bench**, enabling assessment of perception, cognition, and linguistic behavior within a cohesive pipeline.

4.1 Architecture Overview

MIMIC-Chat is composed of three key components: a spatiotemporal video encoder, a task-guided instruction formatter, and a causal language model. This architecture supports both single-choice classification and free-form comment generation under a unified input format.

The model takes a video V and task-specific instruction T as input, and produces the output Y via:

$$Y = \text{LM}([\texttt{VID}], \phi(V)', [\texttt{SEP}], T) \tag{1}$$

Here, $\phi(V)'$ denotes visual tokens extracted and projected from the video, while `[VID]` and `[SEP]` are special tokens marking the multimodal boundaries. This design enables multimodal fusion and instruction-following through a shared interface. It balances open-ended expressiveness with discriminative precision, making it suitable for all tasks in MIMIC-Bench. A detailed architecture diagram is shown in Figure 3.

4.2 Video Encoding and Visual Token Extraction

To convert raw video into structured multimodal input, we adopt a spatiotemporal encoder with a lightweight projection layer, ensuring compatibility with the language model backbone.

Frame Sampling and Preprocessing. We adopt a dual-branch design: the spatial branch uniformly samples 8 frames to capture scene-level cues, while the temporal branch consumes the full frame sequence to preserve fine-grained dynamics and causal context. All frames are resized and center-cropped to a fixed resolution before encoding.

Spatiotemporal Encoding via TimeSformer. The processed frames are input to a spatiotemporal encoding pipeline composed of a spatial encoder (operating on sampled frames) and a temporal encoder (operating on the full sequence), which together capture complementary spatial layouts and temporal dependencies. It outputs a fixed-length token sequence:

$$\phi(V) = \{v_1, v_2, \ldots, v_N\}, \quad v_i \in \mathbb{R}^d \tag{2}$$

MLP-Based Projection. We align visual tokens with the language space via a lightweight MLP:

$$v_i' = \text{MLP}(v_i) \tag{3}$$

For simplicity, Eq. (3) shows a generic projection; in practice, we use distinct spatial and temporal projectors before fusion. The transformed sequence $\phi(V)'$ ensures effective vision-language integration. We distinguish a spatial projector (mapping tokens from the sampled-frame branch) and a temporal projector (mapping tokens from the full-sequence branch). Their outputs are subsequently fused within the language model via gated integration, ensuring complementary spatial–temporal reasoning.

Multimodal Input Construction. We concatenate the visual tokens $\phi(V)'$ with the natural language instruction T, using special tokens `[VID]` and `[SEP]` to delineate modalities. Each token is assigned modality-aware positional embeddings.

Language Model Integration. We employ **InternLM-8B** as the causal language model. It receives the multimodal sequence and generates either a structured classification output or a free-form textual response depending on the task type.

This video encoder module forms the foundation for MIMIC-Chat's high-level reasoning and generation capabilities, providing a consistent visual representation for diverse human-centered video understanding tasks. For a more detailed explanation of the model architecture and training configurations, refer to Appendix, where we provide additional insights on the system architecture and benchmark construction.

4.3 Training Data Construction - Mimic-Data

We build an instruction-tuning dataset named **Mimic-Data** from 150,000+ high-quality videos, each paired with multiple QA examples aligned with MIMIC-Bench tasks. This ensures consistency between training and evaluation in both structure and intent.

For the seven single-choice tasks (CIU, CAM, UIU), each sub-task is framed using a standardized prompt (*e.g.*, *"Based on the video, choose the most appropriate tag/title/comment"*), followed by four candidate options. The correct answer guides the model to align visual-linguistic cues with semantic labels. For the comment generation task, we use a unified prompt per video (*e.g.*, *"Generate a natural and imaginative comment for this video."*) and provide the top-5 most-liked user comments as references. This helps the model learn to express emotional nuance, associative thinking, and stylistic variation.

4.4 Training Objective and Optimization

We use supervised instruction tuning to train MIMIC-Chat on all tasks, optimizing for accurate, human-like responses to video and instructions.

Training Loss. All training samples are cast as question-answer pairs, allowing us to apply a standard language modeling loss:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{|Y|} \log P(y_t \mid X, y_{<t}) \tag{4}$$

This formulation unifies structured classification (*e.g.*, choice "A") and open-ended generation (*e.g*, full comments) under a single decoding objective.

Multi-task Learning. We employ a unified architecture without task-specific heads. Batches from different tasks are randomly sampled, enabling the model to generalize across diverse task types by interpreting the instruction and visual context.

Optimization Strategy. We fine-tune the full model except for the frozen visual backbone. Training leverages standard techniques including mixed-precision computation, dropout, gradient clipping, and label smoothing for improved stability and generalization. This unified objective aligns spatial–temporal semantics with human-centric reasoning goals, supporting both discriminative and generative tasks in MIMIC-Bench.

Parameter-Efficient Tuning. Following (Hu et al., 2022), we integrate LoRA modules into select attention layers of InternLM-8B and jointly train them with the multimodal projector. This approach reduces memory usage while preserving performance. We further examine the contribution of the temporal encoder, spatial/temporal projectors, and LoRA through ablation studies in Appendix.

5 Experimental Evaluation

5.1 Experimental Setup

We evaluate MIMIC-Chat and 24 baselines—including 21 open-source MLLMs and 3 powerful proprietary models—as well as human participants on two key tracks in MIMIC-Bench: (1) **User-like Thinking Tasks**, which cover structured reasoning problems based on metadata; and (2) **User-like Mimicking Tasks**, which assess comment generation.

Implementation Details. For baseline models, we follow each model's official preprocessing strategy (including frame sampling and resolution). For MIMIC-Chat, we adopt the dual-branch design described in Section 4: the spatial branch uniformly samples 8 frames, while the temporal branch processes the full frame sequence. Prompts follow the format defined in Section 4. Subjective evaluation is conducted via a web interface. Three independent annotators judged each comment on two dimensions: (i) whether it was human- or AI-written, and (ii) a realism score between 0 and 5. We summarize key ablation findings here and provide detailed results in the Appendix. Removing the temporal encoder or either projector leads to substantial performance drops, while excluding LoRA causes consistent but smaller degradation.

5.2 User-like Thinking Task Performance

We evaluate all participants on the seven structured reasoning tasks in MIMIC-Bench, spanning three axes: creator intent understanding (CIU), content attribute matching (CAM), and user interaction understanding (UIU). Accuracy is reported per task and averaged across all subtasks. Results are summarized in Table 1.

Human participants achieve an overall accuracy of 73.1%, maintaining strong consistency across all reasoning axes and setting the upper bound on UIU tasks, particularly in Comment Matching

Table 1: Accuracy (%) on the seven reasoning tasks in MIMIC-Bench, covering three cognitive axes: Creator Intent Understanding (TiS: Title Selection, DeS: Description Selection), Content Attribute Matching (TaM: Tag Matching, ToM: Topic Matching, CaM: Category Matching), and User Interaction Understanding (CoM: Comment Matching, CoP: Comment Popularity). **Overall** denotes the average accuracy across all seven tasks. The highest, second-highest, and third-highest scores are highlighted in **purple**, **green**, and **pink**, respectively.

Task Type	**CIU**		**CAM**			**UIU**		**Overall↑**
Models / Tasks	TiS ↑	DeS ↑	TaM ↑	ToM ↑	CaM ↑	CoM ↑	CoP ↑	
VideoChatGPT (Maaz et al., 2024)	27.6	33.1	24.8	23.3	16.7	23.8	22.3	24.1
Video-LLaVA (Lin et al., 2023)	27.0	41.2	68.3	32.4	17.0	24.6	25.8	31.6
VideoChat2 (Li et al., 2024b)	37.7	33.2	30.1	46.1	34.2	32.9	26.7	33.6
LLaVA-NeXT (Li et al., 2024a)	32.7	32.9	50.4	35.3	31.4	25.7	23.5	31.6
VideoLLaMA2 (Cheng et al., 2024)	39.2	38.2	63.8	42.3	36.1	37.7	27.1	39.3
VideoLLaMA3 (Zhang et al., 2025)	78.7	58.0	78.6	82.0	49.5	51.3	28.3	58.7
MiniCPM-V (Yao et al., 2024)	77.6	65.4	77.0	78.6	49.1	55.0	28.9	59.2
MiniCPM-o (Yao et al., 2024)	71.1	60.7	84.8	83.3	50.2	54.2	32.0	59.5
CogVLM2 (Hong et al., 2024)	51.1	37.5	83.6	70.3	55.0	41.9	29.9	50.2
Qwen2-VL (2B) (Wang et al., 2024a)	48.6	51.9	70.8	57.1	47.3	36.2	24.3	44.3
Qwen2-VL (7B) (Wang et al., 2024a)	75.0	75.3	84.7	76.4	33.1	52.2	27.3	57.4
Qwen2.5-VL (3B) (Bai et al., 2025)	78.7	51.2	74.7	85.7	43.9	56.1	29.5	59.0
Qwen2.5-VL (7B) (Wang et al., 2024a)	80.8	54.1	72.6	88.0	43.6	58.1	29.0	59.9
Qwen2.5-VL (72B) (Wang et al., 2024a)	85.6	79.3	79.8	**93.3**	50.6	67.3	33.1	66.7
InternVL2 (2B) (Chen et al., 2024b)	46.1	27.2	67.3	62.7	45.1	33.5	23.8	41.9
InternVL2 (4B) (Chen et al., 2024b)	76.3	52.8	72.8	74.8	45.0	49.8	32.5	56.8
InternVL2 (8B) (Chen et al., 2024b)	84.1	72.9	78.6	87.2	51.1	63.9	30.6	64.4
InternVL2.5 (4B) (Chen et al., 2024a)	79.2	33.3	85.2	84.8	53.7	57.4	28.7	60.7
InternVL2.5 (8B) (Chen et al., 2024a)	83.5	47.6	**86.6**	89.8	50.0	64.0	31.6	64.5
InternVideo2.5 (Wang et al., 2025)	83.2	71.7	**87.5**	90.1	53.5	64.4	32.7	66.3
InternVL3 (78B) (Zhu et al., 2025)	87.4	75.1	80.1	90.5	51.5	70.2	33.3	67.5
ChatGPT-4o (Achiam et al., 2023)	87.9	80.3	83.6	88.7	51.3	70.9	33.5	68.2
Gemini2.5-pro (Comanici et al., 2025)	**92.6**	**89.5**	82.9	**92.3**	**56.1**	**82.9**	43.5	**75.1**
o3 (El-Kishky et al., 2025)	**93.2**	**86.1**	85.7	92.1	55.2	77.4	**45.5**	**74.6**
Human	85.1	77.2	78.7	90.6	**60.0**	**85.9**	**51.1**	73.1
MIMIC-Chat(Ours)	**90.4**	**87.1**	**86.7**	**92.5**	**55.7**	**78.3**	**43.6**	**74.1**

(85.9%) and Comment Popularity (51.1%). **MIMIC-Chat**, our fine-tuned model, attains an overall accuracy of 74.1%, ranking third among all participants—slightly behind two frontier proprietary models (Gemini-2.5-Pro at 75.1% and o3 at 74.6%) but surpassing all open-source baselines, including very large ones such as Qwen2.5-VL-72B (66.7%) and InternVL3-78B (67.5%). MIMIC-Chat achieves top-three performance across nearly all subtasks, with notable strength in creator intent understanding (TiS 90.4%, DeS 87.1%) and content attribute matching (CaM 55.7%).

Despite these gains, UIU remains the most challenging axis: while MIMIC-Chat improves substantially over prior open models, its performance on Comment Popularity (43.6%) still lags far behind human annotators. This highlights the intrinsic difficulty of modeling user-centric semantics and interaction cues.

Overall, the results confirm that human-aligned fine-tuning enables MIMIC-Chat to close the gap with, and in some cases surpass, much larger models, underscoring the importance of task-specific alignment rather than sheer model scale alone.

5.3 User-like Mimicking Task Performance

We evaluate each model's ability to generate human-like comments through a comment simulation task. Table 2 summarizes the results. Here, the first column (*Models*) denotes the comment sources: each row reports evaluation results for comments generated by the corresponding model, while the *Human* row represents real user comments originally collected from video-sharing platforms.

Comment Simulation. For each video, all models generate one comment, which is then evaluated by human annotators. The simulation quality is measured in three ways: (1) the percentage of a model's comments judged as "human"; (2) the distribution of realism scores (0–5); and (3) the mean realism score per model. Each comment is anonymously presented to multiple human evaluators, who are asked to determine whether it was written by a human or an AI, and to rate its human-likeness on a 0–5 scale. Inter-annotator agreement reached 91.95%, calculated as the proportion of samples on which at least two annotators gave consistent judgments. This high level of agreement indicates that the evaluation is stable and less susceptible to individual annotator bias. The "Judged as Human" metric reflects the proportion of a model's comments that were classified as human-written. The "Score@k" distribution indicates the percentage of all comments receiving score k, and "Mean Score" reports the overall average realism score across the model's 4,000 generated comments.

As shown in the table, **MIMIC-Chat significantly outperforms all baseline models** in the key metric of being judged as human (64.24%), more than triple most existing models, and second only to real human comments (87.57%). Its average realism score (2.88) is also notably higher than that of

Table 2: Evaluation of comment simulation across all participants, based on human judgments. The results include human-likeness classification rates, realism score distributions (0–5), and mean scores. The highest, second-highest, and third-highest scores are highlighted in **purple**, **green**, and **pink**, respectively.

Models\Task	Comment Simulation								
	Judged as Human (%) ↑	Judged as AI (%) ↓	Score Distribution(%)						Mean Score ↑
			Score@0 ↓	Score@1 ↓	Score@2	Score@3	Score@4 ↑	Score@5 ↑	
VideoChatGPT (Maaz et al., 2024)	18.65	81.35	50.15	27.52	3.67	7.95	6.27	4.43	1.06
Video-LLaVA (Lin et al., 2023)	6.30	93.70	63.16	28.48	1.88	2.07	2.44	1.97	0.58
VideoChat2 (Li et al., 2024b)	21.02	78.98	53.66	21.54	3.79	9.40	7.70	3.92	1.08
LLaVA-NeXT (Li et al., 2024a)	7.02	92.98	63.20	27.58	2.12	2.93	1.90	2.27	0.60
VideoLLaMA2 (Cheng et al., 2024)	2.70	97.30	75.57	20.37	1.25	0.31	1.56	0.94	0.35
VideoLLaMA3 (Zhang et al., 2025)	4.02	95.98	68.46	24.95	2.47	1.28	1.19	1.65	0.47
MiniCPM-V (Yao et al., 2024)	8.17	91.83	61.78	26.98	2.99	2.64	2.37	3.25	0.67
MiniCPM-o (Yao et al., 2024)	6.85	93.15	66.43	24.60	2.11	1.23	2.81	2.81	0.58
CogVLM2 (Hong et al., 2024)	7.50	92.50	68.20	21.40	2.80	4.80	1.70	1.10	0.54
Qwen2-VL (2B) (Wang et al., 2024a)	6.54	93.46	63.82	27.25	2.39	2.08	2.31	2.16	0.58
Qwen2-VL (7B) (Wang et al., 2024a)	6.99	93.01	62.36	28.25	2.33	2.70	1.95	2.40	0.61
Qwen2.5-VL (3B) (Bai et al., 2025)	14.18	85.82	59.61	23.76	2.38	3.93	4.83	5.49	0.87
Qwen2.5-VL (7B) (Bai et al., 2025)	10.25	89.75	67.76	20.73	1.19	2.60	3.27	4.46	0.66
InternVL2 (2B) (Chen et al., 2024b)	1.03	98.97	80.59	17.66	0.66	0.29	0.29	0.51	0.24
InternVL2 (4B) (Chen et al., 2024b)	3.69	96.31	77.14	17.70	1.47	0.66	1.55	1.47	0.36
InternVL2 (8B) (Chen et al., 2024b)	5.37	94.63	66.09	26.36	2.11	2.18	1.09	2.18	0.52
InternVL2.5 (4B) (Chen et al., 2024a)	16.26	83.74	54.24	26.58	2.93	4.39	5.70	6.16	0.99
InternVL2.5 (8B) (Chen et al., 2024a)	10.90	89.10	59.47	26.54	2.86	2.56	4.21	4.36	0.79
InternVideo2.5 (Wang et al., 2025)	12.67	87.33	55.58	27.60	3.77	4.45	4.45	4.15	0.87
Qwen2.5-VL (72B) (Bai et al., 2025)	26.93	73.07	42.59	22.83	7.65	6.92	9.63	10.40	1.49
InternVL3 (78B) (Zhu et al., 2025)	30.21	69.79	41.23	18.87	9.69	5.48	13.83	10.90	1.65
ChatGPT-4o (Achiam et al., 2023)	40.32	59.68	**28.65**	21.32	9.71	10.91	17.64	11.77	2.03
Gemini2.5-pro (Comanici et al., 2025)	41.87	58.13	29.51	**16.84**	11.78	9.97	18.31	**13.59**	**2.12**
o3 (El-Kishky et al., 2025)	**43.36**	**56.64**	28.68	19.71	8.25	11.33	**19.56**	12.47	2.11
Human	**87.57**	**12.43**	**8.84**	**3.13**	0.49	4.37	**24.59**	**58.58**	**4.08**
MIMIC-Chat (ours)	**64.24**	**35.76**	**7.62**	**15.23**	12.91	26.82	20.20	**17.22**	**2.88**

other models. These results suggest that MIMIC-Chat produces more natural, human-aligned comments in both linguistic style and semantic coherence. These findings highlight that MIMIC-Chat achieves the most human-aligned comment generation performance among all evaluated models.

We attribute this advantage to a fundamental behavioral difference: most baseline models tend to generate literal or content-descriptive comments—such as rephrasing video scenes or restating visual facts—which are easily recognized as AI-written by human annotators. In contrast, MIMIC-Chat demonstrates a stronger capacity for simulating human-like thinking patterns, often exhibiting signs of *divergent, associative, or emotionally reflective reasoning*—traits more characteristic of genuine human responses in open-domain video discussions. Beyond open-source baselines, the inclusion of stronger proprietary models (e.g., ChatGPT-4o, Gemini2.5-pro, and o3) reveals a narrowing gap in human-likeness. These models achieve substantially higher realism scores and human-judgment rates than most open-source MLLMs. Nevertheless, MIMIC-Chat remains the most competitive among open-source systems and still exhibits a notable margin over proprietary counterparts in the "Judged as Human" metric. Taken together, these results indicate that while recent closed-source models are improving rapidly in simulating human-like comments, there remains a sizable gap compared to genuine human responses, underscoring the challenge of fully capturing the subtleties of human reasoning and expression.

To ensure fairness, we also fine-tuned several strong open baselines on MIMIC-Data under identical settings. While their performance improved, they still lagged behind MIMIC-Chat, suggesting that our advantage stems not merely from access to task-aligned data but from the proposed architecture and training strategy. Full results are reported in Appendix.

6 Conclusion

We presented **MIMIC-Data** and **MIMIC-Bench**, a large-scale dataset and benchmark for evaluating human-aligned video understanding, and introduced **MIMIC-Chat**, a model trained to bridge the gap between perception and human-like reasoning. Across 24 competitive MLLMs, results show that while recent systems advance perception, they still fall short in mimicking human thought and expression. MIMIC-Chat narrows this gap, setting a new state of the art among open-source models. We believe this work opens a new direction toward socially and cognitively aligned multimodal intelligence.

Ethics Statement

Our work adheres to the ICLR Code of Ethics. This work introduces MIMIC-Bench, a benchmark designed to evaluate multimodal large language models (MLLMs) on human-aligned video understanding tasks, and MIMIC-Chat, a model fine-tuned for this purpose. All videos and metadata in MIMIC-Data were collected from publicly available sources under permissible use. Personally identifiable or sensitive information was excluded during data collection, and only videos with non-sensitive, non-personal content were retained. The benchmark is intended purely for academic research, and we carefully considered potential risks of misuse, including bias amplification and overfitting to social norms. Our aim is to provide the research community with a rigorous tool for assessing MLLMs' human-aligned capabilities, while promoting responsible stewardship of AI technologies.

Reproducibility Statement

We have made significant efforts to ensure the reproducibility of our work. The main paper details the construction process of MIMIC-Bench, the task definitions, and the evaluation protocols. We describe the design of MIMIC-Chat, including its dual-branch spatial–temporal encoder, training strategy, and optimization settings. Appendix C and D provide further details on model architecture, training configurations, dataset construction, and visualization examples. We plan to release the benchmark data splits, evaluation code, and MIMIC-Chat training scripts upon publication to facilitate replication and follow-up research.

LLM Usage Statement

Large Language Models (LLMs) were used only as part of the evaluation in our benchmark, where 24 baseline models (both open-source and proprietary) were compared against our proposed MIMIC-Chat. LLMs were not used to generate any dataset samples, annotations, or experimental results. All dataset construction, annotation, and analysis were conducted independently by the authors. We used LLMs in a limited capacity for language polishing of the manuscript, but no section of the research design, data collection, or core analysis was generated by LLMs. The authors take full responsibility for the content of this paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024b.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021.

Muhammad Iqbal Hasan Chowdhury, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Hierarchical relational attention for video question answering. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 599–603. IEEE, 2018.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, et al. Competitive programming with large reasoning models. *arXiv preprint arXiv:2502.06807*, 2025.

Gunther Eysenbach et al. The role of chatgpt, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers. *JMIR Medical Education*, 9(1):e46885, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2758–2766, 2017.

Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, May 2024a. URL https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125, 2023.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2025.

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.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024b.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv:2108.10904*, 2021.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.

Qu Yang, Mang Ye, and Bo Du. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*, 2024.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv*, abs/2308.02490, 2023.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. URL https://arxiv.org/abs/2501.13106.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. URL https://arxiv.org/abs/2306.02858.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Ablation Studies

A.1 Ablation Setup

To isolate the contribution of each key component in **MIMIC-Chat**, we perform controlled ablations under a unified training and evaluation pipeline. Unless otherwise specified, all settings strictly follow the main paper (§4–§5): same training data (**MIMIC-Data**), instruction format, loss, optimizer, batch size, number of epochs, inference hardware (A6000 GPUs), and evaluation metrics. For thinking tasks we report accuracy on the seven multiple-choice subtasks; for mimicking we report human-judged comment simulation metrics (*Judged as Human*, Score@k, Mean Score). We do not include source-identification results in the appendix.

Backbone and Inputs (constant across variants). We use the full **dual-branch** video pathway described in §4 as the reference ("Full"): a *spatial branch* that uniformly samples 8 frames for scene-level cues, and a *temporal branch* that processes the full frame sequence for fine-grained dynamics. Visual tokens from both branches are projected by branch-specific MLP projectors and fused into the language model with gated integration. The causal LM is **InternLM-8B** with LoRA modules enabled in the Full configuration.

Ablated Variants. We construct four ablation variants by toggling one component at a time while keeping all other factors identical:

- **w/o LoRA** (*No-LoRA*): Disable all LoRA adapters in InternLM-8B and freeze the LM parameters; only the multimodal projector(s) remain trainable. This tests the role of parameter-efficient language adaptation for instruction alignment.
- **w/o Temporal Encoder** (*No-TempEnc*): Remove the temporal branch (full-sequence processing); the model uses only the spatial branch with 8 uniformly sampled frames. This probes the importance of explicit temporal modeling.
- **w/o Temporal Projector** (*No-TempProj*): Keep both branches but remove the temporal-specific projector; temporal tokens are routed through the spatial projector (shared MLP) before fusion. This examines whether a dedicated temporal projection space is necessary.
- **w/o Spatial Projector** (*No-SpatProj*): Keep both branches but remove the spatial-specific projector; spatial tokens are routed through the temporal projector (shared MLP). This complements the previous variant and tests sensitivity to projector specialization.

Fairness Controls. All variants use the same prompts and decoding settings as the Full model. Frame sampling, resolution, and preprocessing follow §5. When a branch is removed (e.g., *No-TempEnc*), the remaining branch and its projector are unchanged; when a projector is removed (e.g., *No-TempProj/No-SpatProj*), tokens are passed through the remaining projector to keep token dimensionality and downstream interfaces intact. This design ensures that any performance difference can be attributed to the ablated component rather than confounds in optimization or data.

A.2 Quantitative Comparison on Thinking Tasks

We quantify the contribution of each core component on the seven multi-choice tasks (CIU: TiS/DeS; CAM: TaM/ToM/CaM; UIU: CoM/CoP). All settings (data, preprocessing, prompts) are held fixed as in the main experiments; only the indicated module is removed or altered.

Key observations. The ablation results reveal that both temporal modeling and LoRA-based adaptation are indispensable. Removing the *temporal encoder* causes the largest overall drop (74.1 $\rightarrow$ 65.8; -8.3 pp), with especially severe declines in *DeS* (-33.8 pp) and UIU tasks (CoM: -12.9 pp, CoP: -11.0 pp), underscoring the need for end-to-end temporal reasoning. LoRA adaptation is equally critical: ablating it reduces Overall to 66.9 (-7.2 pp), driven by sharp declines on semantics-heavy tasks such as *DeS* (-22.7 pp), *TaM* (-16.3 pp), and both UIU subtasks. The *temporal projector* further contributes complementary gains, as its removal lowers Overall to 67.5 (-6.6 pp) and disproportionately weakens CAM and UIU performance, while the *spatial projector* supports static semantics, with its absence (Overall 68.7; -5.4 pp) most affecting *DeS* (-18.9 pp). Together,

Table 3: Accuracy (%) on the seven structured reasoning tasks after ablating components of MIMIC-Chat. CIU includes Title Selection (TiS) and Description Selection (DeS); CAM includes Tag/Topic/Category Matching (TaM/ToM/CaM); UIU includes Comment Matching (CoM) and Comment Popularity (CoP). **Overall** is the average over all seven tasks. The full model's scores are highlighted in **purple**.

Task Type	**CIU**		**CAM**			**UIU**		**Overall↑**
Models / Tasks	TiS ↑	DeS ↑	TaM ↑	ToM ↑	CaM ↑	CoM ↑	CoP ↑	
MIMIC-Chat (full)	**90.4**	**87.1**	**86.7**	**92.5**	**55.7**	**78.3**	**43.6**	**74.1**
w/o LoRA	88.3	64.4	70.4	89.8	47.3	68.2	34.1	66.9
w/o Temporal Encoder	85.6	53.3	87.5	89.5	51.4	65.4	32.6	65.8
w/o Temporal Projector	89.0	75.4	72.8	91.8	48.6	70.3	34.5	67.5
w/o Spatial Projector	89.3	68.2	84.2	91.7	50.2	72.6	35.1	68.7

these findings highlight that temporal components drive interaction- and context-sensitive reasoning, while LoRA secures creator-intent and textual alignment, and only their integration allows the full model to achieve a balanced advantage across CIU, CAM, and UIU.

A.3 Discussion

The ablation results highlight several broader implications. First, **temporal modeling and LoRA-based language adaptation are complementary**: the former underpins interaction- and context-sensitive reasoning (UIU, CAM), while the latter ensures fine-grained textual alignment (CIU, DeS/TaM). Second, different task categories stress distinct modalities—CIU benefits most from semantic adaptation, whereas UIU requires strong temporal grounding—indicating that balanced multimodal integration is essential for generalizable video understanding. Finally, the consistent superiority of the full model suggests that parameter-efficient language tuning and temporally-aware encoding are not just additive improvements, but jointly critical for bridging the gap between perception-driven reasoning and socially aligned interpretation, reinforcing the design principles behind MIMIC-Chat.

B Fine-tuning Other Models

B.1 Setup

To assess whether the improvements of MIMIC-Chat stem solely from access to MIMIC-Data, we fine-tuned several strong video-language models under identical conditions. Specifically, we selected three representative backbones covering different architectures and scales: **Qwen2.5-VL-7B**, **InternVL2.5-8B**, and **InternVideo2.5-8B**. These models were chosen because of their strong baseline performance and wide adoption in the community.

For fairness, all models were fine-tuned on the same training split of **MIMIC-Data** that was used to train MIMIC-Chat, and evaluated on the official test set of **MIMIC-Bench**. The fine-tuning procedure followed a unified setup:

- **Data.** The full MIMIC-Data training split was used without task-specific resampling. All structured and generative tasks share the same preprocessing pipeline.
- **Optimization.** We employed AdamW optimizer with a learning rate of 2×10^{-5}, cosine decay, and batch size of 128. Training was run for 3 epochs with early stopping on the validation set.
- **LoRA.** For parameter-efficient adaptation, LoRA modules were applied to the language backbone of each model, while vision encoders were kept frozen. This ensured efficiency and comparability across different backbones.
- **Evaluation.** All models were evaluated on the seven structured reasoning tasks (CIU, CAM, UIU) and the mimicking tasks, using accuracy for multi-choice and human-likeness metrics for generative tasks.

Table 4: Accuracy (%) on the seven structured reasoning tasks after fine-tuning existing models on MIMIC-Data. CIU includes Title Selection (TiS) and Description Selection (DeS); CAM includes Tag/Topic/Category Matching (TaM/ToM/CaM); UIU includes Comment Matching (CoM) and Comment Popularity (CoP). **Overall** is the average over all seven tasks. The full model's scores are highlighted in **purple**.

Task Type	**CIU**		**CAM**			**UIU**		**Overall↑**
Models / Tasks	TiS ↑	DeS ↑	TaM ↑	ToM ↑	CaM ↑	CoM ↑	CoP ↑	
Qwen2.5-VL-7B	80.8	54.1	72.6	88.0	43.6	58.1	29.0	59.9
InternVL2.5-8B	83.5	47.6	86.6	89.8	50.0	64.0	31.6	64.5
InternVideo2.5-8B	83.2	71.7	87.5	90.1	53.5	64.4	32.7	66.3
fine-tuned on MIMIC-Data								
Qwen2.5-VL-7B (ft)	85.1	60.2	80.1	89.7	46.8	64.7	32.3	66.4
InternVL2.5-8B (ft)	85.6	53.3	87.5	89.5	51.4	65.4	32.6	65.8
InternVideo2.5-8B (ft)	87.5	76.2	90.3	91.3	51.3	66.9	33.1	68.1
MIMIC-Chat (Ours)	**90.4**	**87.1**	**86.7**	**92.5**	**55.7**	**74.3**	**43.6**	**74.1**

This setup ensures that any observed performance differences are attributable to model architecture and adaptation capacity, rather than data imbalance or training procedure.

B.2 Results

We report the performance of fine-tuned models compared with MIMIC-Chat across both structured reasoning and mimicking tasks.

Key findings. Fine-tuning on MIMIC-Data consistently improves all baseline models, with Qwen2.5-VL-7B gaining from 59.9 to 66.4 overall accuracy and InternVideo2.5-8B rising from 66.3 to 68.1. The strongest gains appear in semantics-heavy tasks such as DeS (e.g., +6.1 for Qwen2.5-VL-7B) and TaM (+7.5), highlighting the value of domain-specific alignment. Nevertheless, MIMIC-Chat remains clearly ahead across nearly all tasks, especially in DeS (87.1 vs. 76.2 for the best fine-tuned baseline) and CoP (43.6 vs. 33.6), confirming that its superior architecture and training design contribute substantially beyond data fine-tuning alone.

B.3 Discussion

The fine-tuning experiments demonstrate that existing MLLMs, when adapted to MIMIC-Data, can indeed improve their performance on user-centric reasoning tasks. Models such as Qwen2.5-VL-7B and InternVideo2.5-8B show consistent gains in both creator-intent understanding (e.g., DeS) and content-attribute matching (e.g., TaM), validating the importance of training data that reflects human communicative patterns. However, the improvements are incremental: even the best fine-tuned baselines remain substantially behind MIMIC-Chat in both overall accuracy and in the most challenging sub-tasks. This suggests that while domain-specific fine-tuning enhances alignment, it cannot substitute for architectural innovations and multi-stage training pipelines explicitly designed for human-like reasoning. In other words, access to the same data is not sufficient—MIMIC-Chat's advantage lies in how it integrates LoRA-based language adaptation, temporal-spatial modeling, and task-specific objectives to achieve balanced and robust performance across all axes of MIMIC-Bench.

C Model Architecture and Implementation Details

In Section 4 of the main paper, we provided a brief overview of the MIMIC-Chat architecture, which integrates a video encoder, an instruction formatter, and a language model into a unified framework. This section supplements that overview by elaborating on implementation-level details and training configurations, including hardware setup, fine-tuning strategies, visual input processing, and optimization techniques.

C.1 Training Environment and Hardware Configuration

All experiments were conducted on a high-performance server equipped with six NVIDIA RTX A6000 GPUs (each with 48 GB memory), using CUDA 12.2 and driver version 535.179. Model training was implemented with PyTorch, and distributed optimization was realized through `torchrun` and DeepSpeed Stage 1, enabling parameter offloading and mixed-precision (bf16) training for enhanced memory and efficiency.

C.2 Fine-tuning Strategy and Module Configuration

MIMIC-Chat adopts a parameter-efficient instruction tuning strategy, updating only key components:

- **Language Model (LLM)**: LoRA modules are injected into the attention sublayers of InternLM2-Chat-8B, leveraging low-rank adaptation to reduce trainable parameters.
- **Freezing Strategy**: Both the vision backbone and LLM backbone are frozen during training, while projection layers and LoRA modules remain trainable.
- **Vision Projection**: Spatial and temporal features are independently projected into the language space via two MLPs to preserve visual-linguistic alignment.

Additionally, bf16 mixed-precision training and gradient checkpointing are enabled to reduce memory usage without compromising performance.

C.3 Video Input Processing and Visual Token Construction

Each video sample undergoes standardized frame sampling and preprocessing:

- **Frame Sampling**: The spatial encoder uses 8 frames uniformly sampled from each video, while the temporal encoder processes the full sequence of frames to capture fine-grained temporal dynamics.
- **Image Processing**: All frames are center-cropped and resized to 448×448 resolution.
- **Feature Encoding**: The spatial and temporal encoders extract static and dynamic information. The spatial encoder processes the 8 uniformly sampled frames, while the temporal encoder consumes the full frame sequence to capture temporal continuity.
- **Projection and Tokenization**: Features from both spatial and temporal encoders are projected into the language model token space to form visual tokens.
- **Input Construction**: Visual tokens and natural language instructions are concatenated, with [VID] and [SEP] tokens denoting modality boundaries.

A dynamic patch control mechanism (up to 6 patches) and thumbnail token injection are introduced to accommodate longer videos and enhance contextual representation.

C.4 Training Configuration and Optimization

To ensure performance and stability, we adopt the following training settings:

- Epochs: 50
- Per-device batch size: 2; Global batch size: 4 (via gradient accumulation)
- Learning rate: 4e-5 with 3% warm-up
- Input resolution: 448×448
- Max dynamic patches: 6
- Optimizer: AdamW with weight decay 0.01
- Scheduler: Cosine decay
- Gradient clipping: enabled

- Max sequence length: 4096 tokens
- Grouped training: samples are grouped by token length to accelerate convergence
- Monitoring: training logs recorded via TensorBoard; best-performing checkpoints and LoRA weights are saved periodically

C.5 Engineering Optimizations for System Robustness

To support long-context, large-scale multimodal training, we introduce several engineering enhancements:

- **Lazy-loading dataset class** for robust video streaming with corrupted frame handling;
- **Custom trainer** with LoRA-only weight saving to facilitate model deployment and ablation analysis;
- **Dynamic image preprocessing** that adapts patch numbers and resolutions on-the-fly to control memory usage;
- **Multi-task training support**, enabling unified classification and generation under instruction-based prompts.

D Benchmark Construction and Visualization Examples

In Section 3 of the main paper, we outlined the construction of MIMIC-Bench and the motivation for its design. This section provides additional implementation details regarding how the 4,000 benchmark videos were selected and scored, including the criteria used to ensure their human-centric relevance and linguistic richness. It also supplements the dataset composition and preparation steps that underpin our evaluation tasks.

D.1 Selection and Scoring Criteria

To ensure that the benchmark accurately reflects human-style interpretation and communicative behavior, we curated 4,000 videos from the larger MIMIC-Data pool of 150,000+ user-shared videos. The selection process involved a multi-stage filtering pipeline:

(1) Engagement Scoring. Each video was assigned a composite engagement score to measure real-world user interaction. The score combines the log-normalized values of like count, favorite count, share count, and comment count, computed as:

$$\text{Engagement Score} = \alpha \cdot \log(\mathit{Like}) + \beta \cdot \log(\mathit{Favorite}) + \gamma \cdot \log(\mathit{Share}) + \delta \cdot \log(\mathit{Comment}) \quad (5)$$

We set $\alpha = 1.0$, $\beta = 0.8$, $\gamma = 0.5$, and $\delta = 1.2$ to place greater emphasis on comments, which better reflect human intent and understanding.

(2) Metadata Integrity. After sorting by engagement score, we retained only those videos with complete metadata fields, including title, description, tags, and topic. We further ensured that each video contains at least five unique, high-quality user comments and is free from decoding errors or anomalously short durations.

(3) Semantic Coverage and Diversity. To ensure diverse coverage across topics and expression styles, we adopted the following constraints:

- Top 2% videos from TikTok and top 5% from YouTube were selected.
- The selected pool spans 8 major categories (e.g., lifestyle, travel, beauty) and 20+ subcategories.
- The distribution of comment types was controlled to include exclamatory, inquisitive, associative, and ironic styles.

This multi-dimensional curation strategy ensures that MIMIC-Bench captures both high user engagement and rich human-centered semantics, laying a robust foundation for downstream evaluation of multimodal models on user-aligned reasoning and mimicking capabilities.

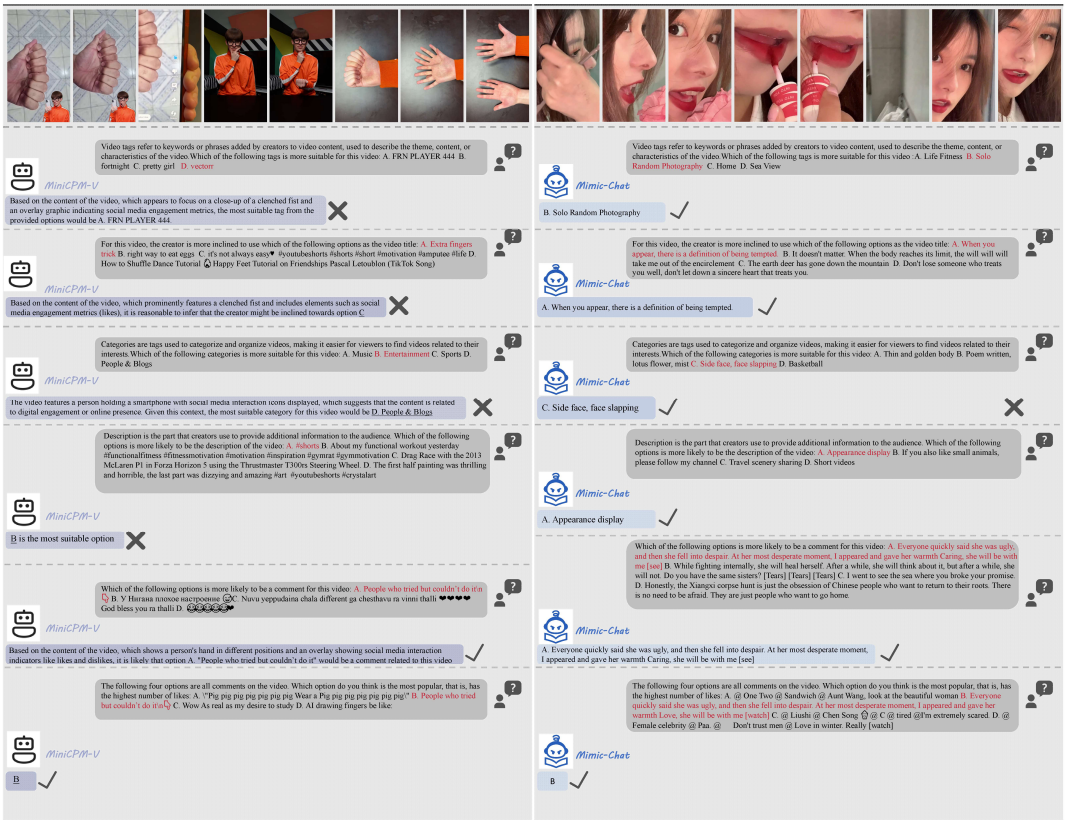

Figure 4: Presentation of the responses to the **MiniCPM-V** and **MIMIC-Chat (Ours)** part of the task. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.

D.2 Qualitative Examples of Model Responses

To better illustrate the performance of different multimodal large language models (MLLMs) on MIMIC-Bench tasks, we present a series of representative model response examples in this supplementary material. These examples cover a variety of tasks such as title selection, tag matching, comment imitation, demonstrating each model's ability to interpret real-world user videos and generate human-aligned responses.

Each example includes the following components:

- **Task input**: the multimodal metadata associated with the video, along with the task prompt;
- **Model-generated responses**: the outputs from a set of baseline MLLMs, as well as our proposed MIMIC-Chat model;
- **Ground-truth or reference answers**: provided for comparison to evaluate model correctness or human-likeness.

As shown in Figures 4, 5, and 6, the visualized outputs display the input prompts, the model predictions, and whether the generated results match the expected answers. These examples qualitatively complement the quantitative results in the main paper, highlighting each model's strengths and weaknesses across tasks involving higher-level reasoning, creative intent recognition, and user interaction interpretation.

We hope these examples will deepen understanding of the challenges posed by MIMIC-Bench and inspire the development of more human-aligned multimodal systems.

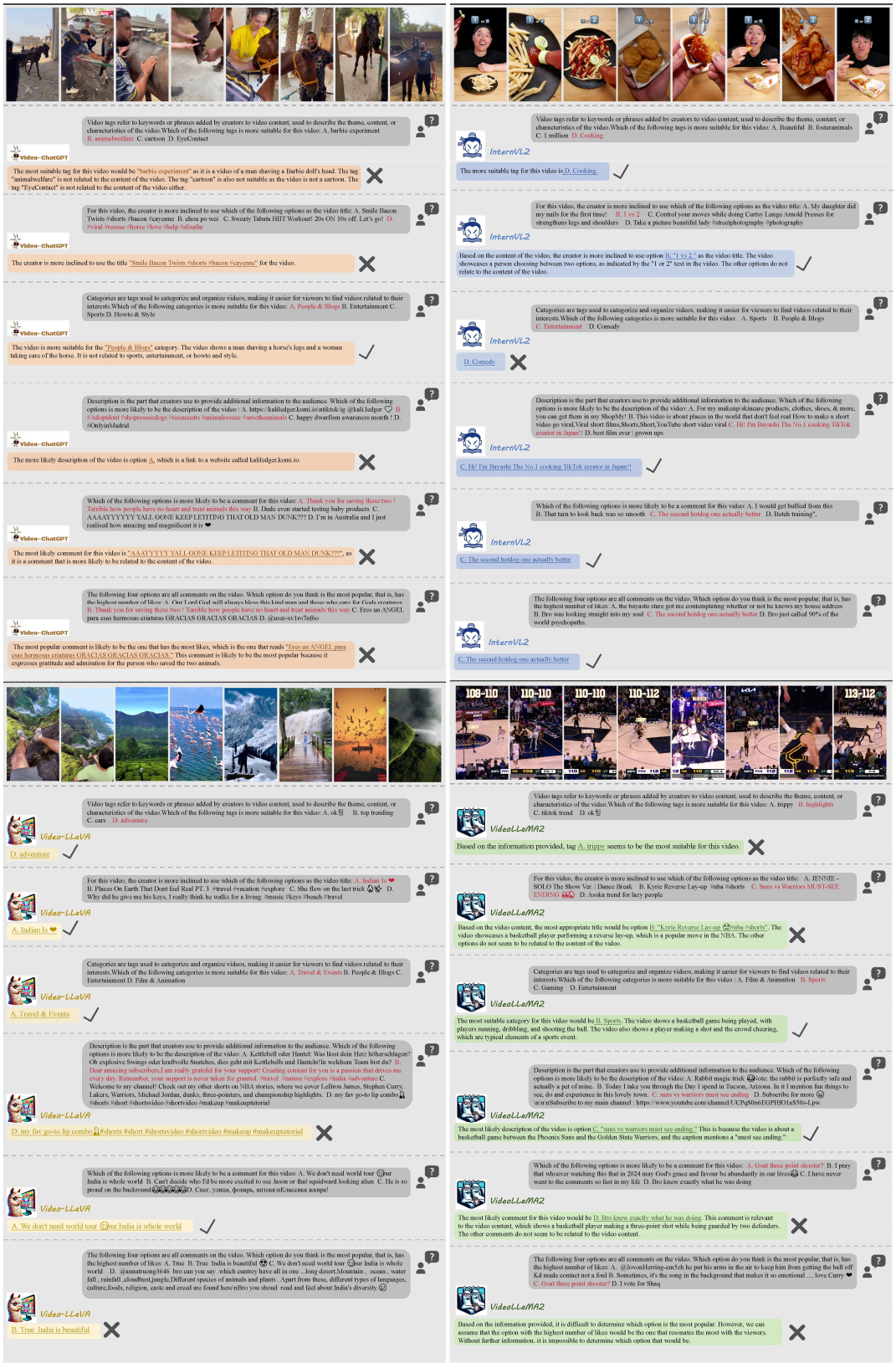

Figure 5: Presentation of the responses to the **Video-ChatGPT**, **InternVL2**, **Video-LLaVA**, and **VideoLLaMA2** part of the task. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.

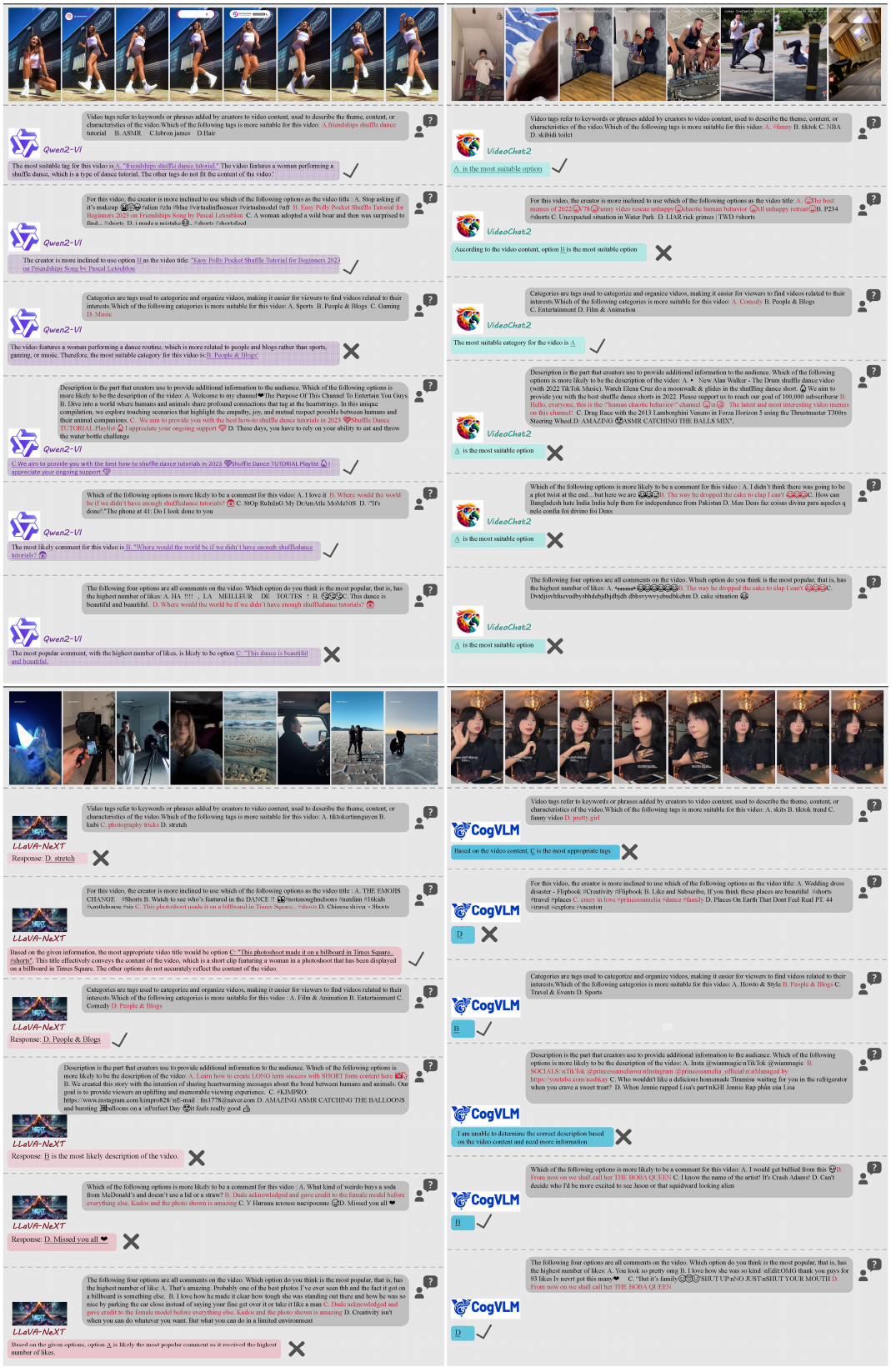

Figure 6: Presentation of the responses to the **Qwen2-VL**, **VideoChat2**, **LLaVA-NeXT**, and **CogVLM2** part of the task. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.

D.3 Extended Description of MIMIC-Data

MIMIC-Data is the foundational dataset for constructing all tasks in MIMIC-Bench. It contains over 150,000 user-generated short videos collected from multiple public video-sharing platforms. While the main paper already outlines the high-level data pipeline and task mappings, this section provides additional implementation details regarding its structure and usage.

Each data sample is stored in structured JSONL format, with fields including `video_path`, `title`, `description`, `tags`, `topic`, and a list of user `comments`. Every video is associated with at least five real user comments. All text fields are pre-cleaned by removing duplicates, empty or meaningless entries, and normalizing punctuation and encoding formats to ensure natural linguistic quality.

During task construction, each video may yield multiple question–answer pairs depending on the completeness of its metadata and the number of available comments. All training prompts are formulated in a unified instruction-following format, where the task type and target field are explicitly encoded (e.g., "Please select the most likely title," or "Which comment is most popular?").

MIMIC-Data is also structurally well-suited for supporting the full range of tasks in MIMIC-Bench. All evaluation samples are derived directly from original metadata fields without requiring additional human annotations. In particular, for the comment imitation tasks, we select the top five most-liked user comments per video and control the stylistic diversity of samples across expressive, associative, declarative, and rhetorical styles to better support modeling of human-like language behavior.

In future releases, we plan to extend MIMIC-Data with multilingual versions and enhanced semantic annotations to support broader research in multimodal reasoning and generation.