

SUPPLEMENTARY MATERIALS FOR MIMIC-BENCH: EXPLORING THE USER-LIKE THINKING AND MIM- ICKING CAPABILITIES OF MULTIMODAL LARGE LAN- GUAGE MODELS

Jiajie Teng¹, Huiyu Duan^{1,*}, Sijing Wu¹, Jiarui Wang¹, Xilei Zhu¹, Jianing Jin¹,
Wei Shen¹, Xiongkuo Min¹, Guangtao Zhai¹

¹Shanghai Jiao Tong University

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.

*Corresponding author.

Table 1: 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
<i>w/o LoRA</i>	88.3	64.4	70.4	89.8	47.3	68.2	34.1	66.9
<i>w/o Temporal Encoder</i>	85.6	53.3	87.5	89.5	51.4	65.4	32.6	65.8
<i>w/o Temporal Projector</i>	89.0	75.4	72.8	91.8	48.6	70.3	34.5	67.5
<i>w/o Spatial Projector</i>	89.3	68.2	84.2	91.7	50.2	72.6	35.1	68.7

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 → 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, 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:

Table 2: 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
<i>fine-tuned on MIMIC-Data</i>								
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

- **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.

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:

$$\text{Engagement Score} = \alpha \cdot \log(\text{Like}) + \beta \cdot \log(\text{Favorite}) + \gamma \cdot \log(\text{Share}) + \delta \cdot \log(\text{Comment}) \quad (1)$$

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.

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 1, 2, and 3, 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.

D.3 DIFFICULTY CHARACTERISTICS OF USER-LIKE THINKING TASKS

The user-like thinking tasks in MIMIC-Bench are deliberately designed to span multiple levels of cognitive complexity. As reported in the main paper, the majority of tasks remain challenging for current multimodal large language models (MLLMs), with substantial gaps to human performance on many user-oriented reasoning problems. Although a small subset of tasks may appear relatively simple from a human perspective, they play an important role in comprehensively probing different aspects of user-like thinking and providing coverage over realistic interaction scenarios.

At a coarse level, some user-like thinking tasks, such as *Title Selection* (TiS), involve relatively direct alignment between the video content, its associated metadata, and the candidate options. Even in these seemingly straightforward cases, models must still capture how human users abstract and summarize video content into concise, platform-style titles. In contrast, higher-level tasks, including *Category Matching* (CaM), *Comment Matching* (CoM), and *Comment Popularity* (CoP), demand reasoning that goes well beyond surface-level cues. These tasks require models to infer implicit attributes, thematic foci, emotional or stylistic undertones, and preference patterns that govern which comments are considered relevant, engaging, or popular by real users.

Importantly, the difficulty of user-like thinking tasks does not primarily stem from visual complexity, but from the need to bridge perceptual and textual information with human-centric cognition. For instance, solving CoM and CoP tasks often hinges on understanding why a particular comment

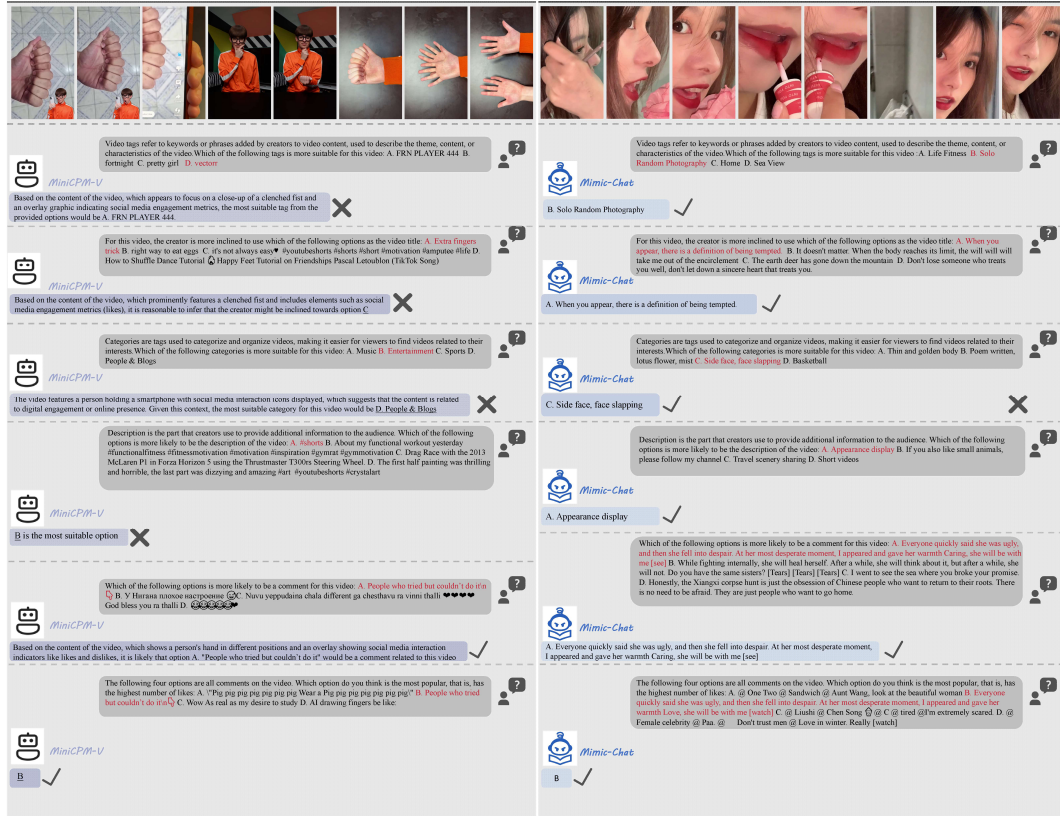


Figure 1: Presentation of the responses to the **MiniCPM-V** and **MIMIC-Chat (Ours)** part of the task. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.

resonates with viewers, how it reflects shared background knowledge or community norms, and how subtle shifts in tone or stance can alter perceived appropriateness or popularity. Such challenges expose fundamental gaps between current MLLMs and the way humans interpret, evaluate, and respond to short-form video content in everyday platforms.

To concretely illustrate these characteristics, Figure 4 presents a set of representative and challenging examples drawn from the user-like thinking tasks, including TiS, CaM, CoM, and CoP. In these cases, successful task completion depends on implicit reasoning about content attributes, expressive intent, or user perspective, rather than merely matching superficial keywords or literal descriptions. Together, these examples highlight that, even when the task interface may look simple, the underlying cognitive demands in MIMIC-Bench are substantial for current MLLMs, and that covering both simpler and more complex user-like thinking scenarios is necessary for a holistic evaluation.

D.4 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.

Figure 2 displays a grid of 16 examples, each showing a video thumbnail, a question, and model responses from four different models: Video-ChatGPT, InternVL2, Video-LLaVA, and VideoLLaMA2. The questions are related to video content, and the model responses provide answers or commentary. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.

Example 1 (Top Left): Video thumbnail shows a person shaving a horse's legs. Question: "The creator is more inclined to use the title 'Smile Bacon Twists rubs bacon xavene' for the video." Model responses: Video-ChatGPT (Incorrect), InternVL2 (Correct), Video-LLaVA (Correct), VideoLLaMA2 (Correct).

Example 2 (Top Right): Video thumbnail shows a person cooking. Question: "The more suitable tag for this video is D. Cooking." Model responses: Video-ChatGPT (Correct), InternVL2 (Correct), Video-LLaVA (Correct), VideoLLaMA2 (Correct).

Example 3 (Middle Left): Video thumbnail shows a person shaving a horse's legs. Question: "The creator is more inclined to use the title 'Smile Bacon Twists rubs bacon xavene' for the video." Model responses: Video-ChatGPT (Correct), InternVL2 (Correct), Video-LLaVA (Correct), VideoLLaMA2 (Correct).

Example 4 (Middle Right): Video thumbnail shows a person cooking. Question: "The more suitable tag for this video is D. Cooking." Model responses: Video-ChatGPT (Correct), InternVL2 (Correct), Video-LLaVA (Correct), VideoLLaMA2 (Correct).

Example 5 (Bottom Left): Video thumbnail shows a person shaving a horse's legs. Question: "The creator is more inclined to use the title 'Smile Bacon Twists rubs bacon xavene' for the video." Model responses: Video-ChatGPT (Correct), InternVL2 (Correct), Video-LLaVA (Correct), VideoLLaMA2 (Correct).

Example 6 (Bottom Right): Video thumbnail shows a person cooking. Question: "The more suitable tag for this video is D. Cooking." Model responses: Video-ChatGPT (Correct), InternVL2 (Correct), Video-LLaVA (Correct), VideoLLaMA2 (Correct).

Figure 2: Presentation of the responses to the Video-ChatGPT, InternVL2, Video-LLaVA, and VideoLLaMA2 all part of the task. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.

Figure 3 displays the responses of four video understanding models (Qwen2-VL, VideoChat2, LLaVA-NeXT, and CogVLM2) to a series of video-based questions. Each row corresponds to a specific video, and each column shows the model's response to a question about that video. The questions are designed to test the models' ability to understand video content, identify key elements, and provide relevant information.

The questions and model responses are as follows:

- Video 1 (Dance Tutorial):**
 - Qwen2-VL: "Video tags refer to keywords or phrases added by creators to video content, used to describe the theme, content, or characteristics of the video. Which of the following tags is more suitable for this video: A. #friendships-shuffle-dance tutorial B. ASMR C. #droner janes D. Hair" (Correct: A)
 - VideoChat2: "Video tags refer to keywords or phrases added by creators to video content, used to describe the theme, content, or characteristics of the video. Which of the following tags is more suitable for this video: A. #friend B. #droner C. #droner D. #droner" (Correct: A)
 - LLaVA-NeXT: "Video tags refer to keywords or phrases added by creators to video content, used to describe the theme, content, or characteristics of the video. Which of the following tags is more suitable for this video: A. #friendships-shuffle-dance tutorial B. ASMR C. #droner janes D. Hair" (Correct: A)
 - CogVLM: "Video tags refer to keywords or phrases added by creators to video content, used to describe the theme, content, or characteristics of the video. Which of the following tags is more suitable for this video: A. #friend B. #droner C. #droner D. #droner" (Correct: A)
- Video 2 (Dance Tutorial):**
 - Qwen2-VL: "The creator is more inclined to use one of the following options as the video title: 'Tara Polly Pocket Shuffle Tutorial for Beginners 2022' on Friendships Song by Pascal Lefebvre" (Correct: A)
 - VideoChat2: "For this video, the creator is more inclined to use which of the following options as the video title: A. To best of my knowledge, this is the 'Summer dance tutorial' channel. B. To best of my knowledge, this is the 'Summer dance tutorial' channel. C. To best of my knowledge, this is the 'Summer dance tutorial' channel. D. To best of my knowledge, this is the 'Summer dance tutorial' channel" (Correct: A)
 - LLaVA-NeXT: "For this video, the creator is more inclined to use which of the following options as the video title: A. To best of my knowledge, this is the 'Summer dance tutorial' channel. B. To best of my knowledge, this is the 'Summer dance tutorial' channel. C. To best of my knowledge, this is the 'Summer dance tutorial' channel. D. To best of my knowledge, this is the 'Summer dance tutorial' channel" (Correct: A)
 - CogVLM: "For this video, the creator is more inclined to use which of the following options as the video title: A. To best of my knowledge, this is the 'Summer dance tutorial' channel. B. To best of my knowledge, this is the 'Summer dance tutorial' channel. C. To best of my knowledge, this is the 'Summer dance tutorial' channel. D. To best of my knowledge, this is the 'Summer dance tutorial' channel" (Correct: A)
- Video 3 (Dance Tutorial):**
 - Qwen2-VL: "Categories are tags used to categorize and organize videos, making it easier for viewers to find videos related to their interests. Which of the following categories is more suitable for this video: A. Sports B. People & Blogs C. Gaming D. Music" (Correct: B)
 - VideoChat2: "Categories are tags used to categorize and organize videos, making it easier for viewers to find videos related to their interests. Which of the following categories is more suitable for this video: A. Comedy B. People & Blogs C. Entertainment D. Film & Animation" (Correct: B)
 - LLaVA-NeXT: "Categories are tags used to categorize and organize videos, making it easier for viewers to find videos related to their interests. Which of the following categories is more suitable for this video: A. Film & Animation B. Entertainment C. Comedy D. People & Blogs" (Correct: B)
 - CogVLM: "Categories are tags used to categorize and organize videos, making it easier for viewers to find videos related to their interests. Which of the following categories is more suitable for this video: A. Howto & Style B. People & Blogs C. Travel & Events D. Sports" (Correct: B)
- Video 4 (Dance Tutorial):**
 - Qwen2-VL: "Description is the part that creators use to provide additional information to the audience. Which of the following options is more likely to be the description of the video: A. Welcome to my channel! B. The Purpose Of This Channel Is Entertain You Guys B. Dive into a world where humans and animals share profound connections that tug at the heartstrings. In this unique compilation, we explore touching scenarios that highlight the empathy, joy, and mutual respect possible between humans and their animal companions. C. We aim to provide you with the best of the best shuffle dance tutorials in 2022! D. Shuffle Dance TUTORIAL Playlist! I appreciate your ongoing support! D. These days, you have to rely on your ability to eat and throw the water bottle challenge" (Correct: C)
 - VideoChat2: "Description is the part that creators use to provide additional information to the audience. Which of the following options is more likely to be the description of the video: A. New Alan Walker & The Drum shuffle dance video (with 2022 TikTok Music). Watch Elena Cruz do a moonwalk & glide in the shuffle dance short. Q We aim to provide you with the best shuffle dance shorts in 2022. Please support us to reach our goal of 100,000 subscribers! B. Hello, everyone, this is the 'Summer dance tutorial' channel. C. To best of my knowledge, this is the 'Summer dance tutorial' channel. D. To best of my knowledge, this is the 'Summer dance tutorial' channel" (Correct: C)
 - LLaVA-NeXT: "Description is the part that creators use to provide additional information to the audience. Which of the following options is more likely to be the description of the video: A. Welcome to my channel! B. The Purpose Of This Channel Is Entertain You Guys B. Dive into a world where humans and animals share profound connections that tug at the heartstrings. In this unique compilation, we explore touching scenarios that highlight the empathy, joy, and mutual respect possible between humans and their animal companions. C. We aim to provide you with the best of the best shuffle dance tutorials in 2022! D. Shuffle Dance TUTORIAL Playlist! I appreciate your ongoing support! D. These days, you have to rely on your ability to eat and throw the water bottle challenge" (Correct: C)
 - CogVLM: "Description is the part that creators use to provide additional information to the audience. Which of the following options is more likely to be the description of the video: A. Welcome to my channel! B. The Purpose Of This Channel Is Entertain You Guys B. Dive into a world where humans and animals share profound connections that tug at the heartstrings. In this unique compilation, we explore touching scenarios that highlight the empathy, joy, and mutual respect possible between humans and their animal companions. C. We aim to provide you with the best of the best shuffle dance tutorials in 2022! D. Shuffle Dance TUTORIAL Playlist! I appreciate your ongoing support! D. These days, you have to rely on your ability to eat and throw the water bottle challenge" (Correct: C)
- Video 5 (Dance Tutorial):**
 - Qwen2-VL: "Which of the following options is more likely to be a comment for this video: A. I love it! B. Where would the world be if we didn't have enough shuffle dance tutorials? C. Stop! Rubbin' My Deft Affle ModleSIS D. 'Tis done! The phone at 41: Do I look down to you" (Correct: B)
 - VideoChat2: "Which of the following options is more likely to be a comment for this video: A. I didn't think there was going to be a plot twist at the end... but here we are! B. The way he dropped the cake to clap I can't! C. How can Bangladesh help India help them for independence from Pakistan D. Man Deus the coias divina para apopleja q vide confia for divine for Dene" (Correct: B)
 - LLaVA-NeXT: "Which of the following options is more likely to be a comment for this video: A. I love it! B. Where would the world be if we didn't have enough shuffle dance tutorials? C. Stop! Rubbin' My Deft Affle ModleSIS D. 'Tis done! The phone at 41: Do I look down to you" (Correct: B)
 - CogVLM: "Which of the following options is more likely to be a comment for this video: A. I didn't think there was going to be a plot twist at the end... but here we are! B. The way he dropped the cake to clap I can't! C. How can Bangladesh help India help them for independence from Pakistan D. Man Deus the coias divina para apopleja q vide confia for divine for Dene" (Correct: B)
- Video 6 (Dance Tutorial):**
 - Qwen2-VL: "The most popular comment, with the highest number of likes, is likely to be option C. 'This dance is beautiful and beautiful'" (Correct: C)
 - VideoChat2: "The most popular comment, with the highest number of likes, is likely to be option C. 'This dance is beautiful and beautiful'" (Correct: C)
 - LLaVA-NeXT: "The most popular comment, with the highest number of likes, is likely to be option C. 'This dance is beautiful and beautiful'" (Correct: C)
 - CogVLM: "The most popular comment, with the highest number of likes, is likely to be option C. 'This dance is beautiful and beautiful'" (Correct: C)
- Video 7 (Dance Tutorial):**
 - Qwen2-VL: "Video tags refer to keywords or phrases added by creators to video content, used to describe the theme, content, or characteristics of the video. Which of the following tags is more suitable for this video: A. #tiktoktopography B. kubi C. #photography risks D. stretch" (Correct: C)
 - VideoChat2: "Video tags refer to keywords or phrases added by creators to video content, used to describe the theme, content, or characteristics of the video. Which of the following tags is more suitable for this video: A. #tiktoktopography B. kubi C. #photography risks D. stretch" (Correct: C)
 - LLaVA-NeXT: "Video tags refer to keywords or phrases added by creators to video content, used to describe the theme, content, or characteristics of the video. Which of the following tags is more suitable for this video: A. #tiktoktopography B. kubi C. #photography risks D. stretch" (Correct: C)
 - CogVLM: "Video tags refer to keywords or phrases added by creators to video content, used to describe the theme, content, or characteristics of the video. Which of the following tags is more suitable for this video: A. #tiktoktopography B. kubi C. #photography risks D. stretch" (Correct: C)
- Video 8 (Dance Tutorial):**
 - Qwen2-VL: "For this video, the creator is more inclined to use which of the following options as the video title: A. THE EMOIS CHANCE 'Shout B. Wants to see who's fatter in the DANCE! C. #newshoesdances thedram H4kds #wonderlous this C. This photoshoot made it on a billboard in Times Square. D. Chance dleves - Shouts" (Correct: C)
 - VideoChat2: "For this video, the creator is more inclined to use which of the following options as the video title: A. Wedding dress disaster - Flipbook C. Creativity Flipbook B. Like and Subscribe: If you think these places are beautiful! Shouts 'napped C. create in love-princessmika video. 'simly D. Places On Earth That Don't Feel Real! IT. 44 'napped recreation" (Correct: C)
 - LLaVA-NeXT: "For this video, the creator is more inclined to use which of the following options as the video title: A. THE EMOIS CHANCE 'Shout B. Wants to see who's fatter in the DANCE! C. #newshoesdances thedram H4kds #wonderlous this C. This photoshoot made it on a billboard in Times Square. D. Chance dleves - Shouts" (Correct: C)
 - CogVLM: "For this video, the creator is more inclined to use which of the following options as the video title: A. THE EMOIS CHANCE 'Shout B. Wants to see who's fatter in the DANCE! C. #newshoesdances thedram H4kds #wonderlous this C. This photoshoot made it on a billboard in Times Square. D. Chance dleves - Shouts" (Correct: C)
- Video 9 (Dance Tutorial):**
 - Qwen2-VL: "Based on the given information, the most appropriate video title would be option C. 'This photoshoot made it on a billboard in Times Square. #shirts' This title effectively conveys the content of the video, which is a short clip featuring a woman in a photoshoot that has been displayed on a billboard in Times Square. The other options do not accurately reflect the content of the video." (Correct: C)
 - VideoChat2: "Based on the given information, the most appropriate video title would be option C. 'This photoshoot made it on a billboard in Times Square. #shirts' This title effectively conveys the content of the video, which is a short clip featuring a woman in a photoshoot that has been displayed on a billboard in Times Square. The other options do not accurately reflect the content of the video." (Correct: C)
 - LLaVA-NeXT: "Based on the given information, the most appropriate video title would be option C. 'This photoshoot made it on a billboard in Times Square. #shirts' This title effectively conveys the content of the video, which is a short clip featuring a woman in a photoshoot that has been displayed on a billboard in Times Square. The other options do not accurately reflect the content of the video." (Correct: C)
 - CogVLM: "Based on the given information, the most appropriate video title would be option C. 'This photoshoot made it on a billboard in Times Square. #shirts' This title effectively conveys the content of the video, which is a short clip featuring a woman in a photoshoot that has been displayed on a billboard in Times Square. The other options do not accurately reflect the content of the video." (Correct: C)
- Video 10 (Dance Tutorial):**
 - Qwen2-VL: "Categories are tags used to categorize and organize videos, making it easier for viewers to find videos related to their interests. Which of the following categories is more suitable for this video: A. Film & Animation B. Entertainment C. Comedy D. People & Blogs" (Correct: B)
 - VideoChat2: "Categories are tags used to categorize and organize videos, making it easier for viewers to find videos related to their interests. Which of the following categories is more suitable for this video: A. Howto & Style B. People & Blogs C. Travel & Events D. Sports" (Correct: B)
 - LLaVA-NeXT: "Categories are tags used to categorize and organize videos, making it easier for viewers to find videos related to their interests. Which of the following categories is more suitable for this video: A. Film & Animation B. Entertainment C. Comedy D. People & Blogs" (Correct: B)
 - CogVLM: "Categories are tags used to categorize and organize videos, making it easier for viewers to find videos related to their interests. Which of the following categories is more suitable for this video: A. Howto & Style B. People & Blogs C. Travel & Events D. Sports" (Correct: B)
- Video 11 (Dance Tutorial):**
 - Qwen2-VL: "Description is the part that creators use to provide additional information to the audience. Which of the following options is more likely to be the description of the video: A. I can't help it, I'm a little bit of a nerd! B. I can't help it, I'm a little bit of a nerd! C. I can't help it, I'm a little bit of a nerd! D. I can't help it, I'm a little bit of a nerd!" (Correct: C)
 - VideoChat2: "Description is the part that creators use to provide additional information to the audience. Which of the following options is more likely to be the description of the video: A. I can't help it, I'm a little bit of a nerd! B. I can't help it, I'm a little bit of a nerd! C. I can't help it, I'm a little bit of a nerd! D. I can't help it, I'm a little bit of a nerd!" (Correct: C)
 - LLaVA-NeXT: "Description is the part that creators use to provide additional information to the audience. Which of the following options is more likely to be the description of the video: A. I can't help it, I'm a little bit of a nerd! B. I can't help it, I'm a little bit of a nerd! C. I can't help it, I'm a little bit of a nerd! D. I can't help it, I'm a little bit of a nerd!" (Correct: C)
 - CogVLM: "Description is the part that creators use to provide additional information to the audience. Which of the following options is more likely to be the description of the video: A. I can't help it, I'm a little bit of a nerd! B. I can't help it, I'm a little bit of a nerd! C. I can't help it, I'm a little bit of a nerd! D. I can't help it, I'm a little bit of a nerd!" (Correct: C)
- Video 12 (Dance Tutorial):**
 - Qwen2-VL: "Which of the following options is more likely to be a comment for this video: A. What kind of words boys a soda from McDonald's and deuce? You a fat or a nerd? B. Dude acknowledged and gave credit to the female model before everything else. Kudos and the photo shoots is amazing C. Y Hiram inxose microphone QD. Missed you all! D. Missed you all!" (Correct: C)
 - VideoChat2: "Which of the following options is more likely to be a comment for this video: A. What kind of words boys a soda from McDonald's and deuce? You a fat or a nerd? B. Dude acknowledged and gave credit to the female model before everything else. Kudos and the photo shoots is amazing C. Y Hiram inxose microphone QD. Missed you all! D. Missed you all!" (Correct: C)
 - LLaVA-NeXT: "Which of the following options is more likely to be a comment for this video: A. What kind of words boys a soda from McDonald's and deuce? You a fat or a nerd? B. Dude acknowledged and gave credit to the female model before everything else. Kudos and the photo shoots is amazing C. Y Hiram inxose microphone QD. Missed you all! D. Missed you all!" (Correct: C)
 - CogVLM: "Which of the following options is more likely to be a comment for this video: A. What kind of words boys a soda from McDonald's and deuce? You a fat or a nerd? B. Dude acknowledged and gave credit to the female model before everything else. Kudos and the photo shoots is amazing C. Y Hiram inxose microphone QD. Missed you all! D. Missed you all!" (Correct: C)
- Video 13 (Dance Tutorial):**
 - Qwen2-VL: "The following four options are all comments on the video. Which option do you think is the most popular, that is, has the highest number of likes: A. You look so pretty omg B. I love how she was so kind and she said thank you guys for 93 likes I've never put this many C. 'But it's family' SHUT UP! NO JUST ASHSH YOUR MOUTH D. From now on we shall call her THE DOBA QUEEN" (Correct: B)
 - VideoChat2: "The following four options are all comments on the video. Which option do you think is the most popular, that is, has the highest number of likes: A. You look so pretty omg B. I love how she was so kind and she said thank you guys for 93 likes I've never put this many C. 'But it's family' SHUT UP! NO JUST ASHSH YOUR MOUTH D. From now on we shall call her THE DOBA QUEEN" (Correct: B)
 - LLaVA-NeXT: "The following four options are all comments on the video. Which option do you think is the most popular, that is, has the highest number of likes: A. You look so pretty omg B. I love how she was so kind and she said thank you guys for 93 likes I've never put this many C. 'But it's family' SHUT UP! NO JUST ASHSH YOUR MOUTH D. From now on we shall call her THE DOBA QUEEN" (Correct: B)
 - CogVLM: "The following four options are all comments on the video. Which option do you think is the most popular, that is, has the highest number of likes: A. You look so pretty omg B. I love how she was so kind and she said thank you guys for 93 likes I've never put this many C. 'But it's family' SHUT UP! NO JUST ASHSH YOUR MOUTH D. From now on we shall call her THE DOBA QUEEN" (Correct: B)

Figure 3: 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.

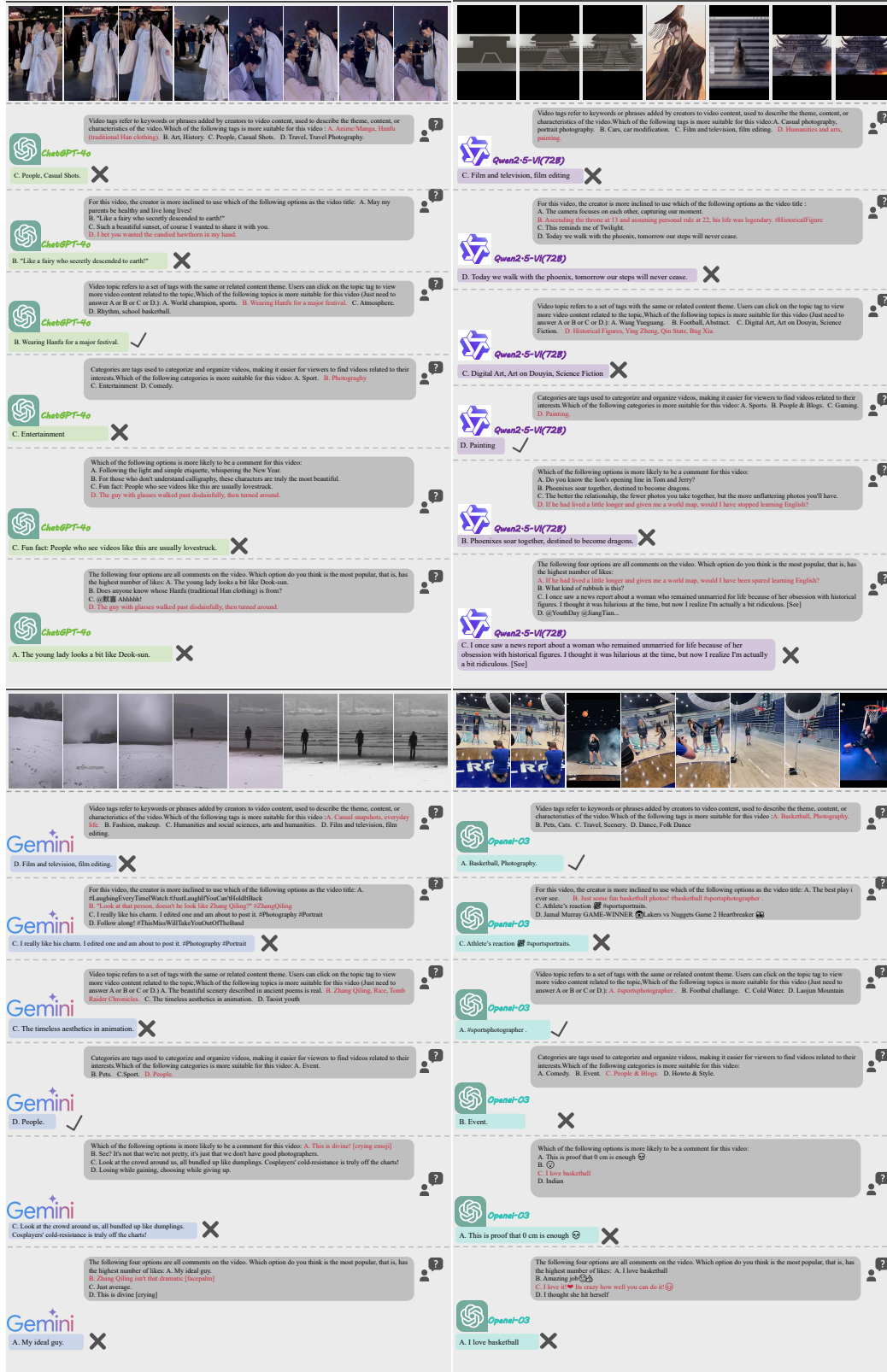


Figure 4: Illustration of the responses from the ChatGPT-4o, Qwen2.5-VL(72B), Gemini2.5-pro, and Openai-o3 models on user-like reasoning tasks. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.

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.

D.5 COMMENT MIMICKING EVALUATION PROTOCOL

The goal of the user-like comment mimicking task is to assess whether models can generate comments that exhibit human-like expressive characteristics commonly observed on social media platforms. Such expressions often involve emotional nuance, metaphor, sarcasm, cultural context, and implicit cross-context associations. At present, there is no stable or widely accepted automatic metric that can reliably evaluate these open-ended, human-centered properties. Therefore, human evaluation is not a limitation of our design, but a necessary component, consistent with common practice in dialogue evaluation, story generation, and multimodal language generation benchmarks.

Evaluation Dimensions. The evaluation focuses on the human-likeness of generated comments and adopts two complementary dimensions: (1) *Comment Origin Discrimination*, where annotators judge whether a comment is more likely written by a human or an AI model; and (2) *Human-likeness Scoring*, where annotators assign a score from 0 to 5 reflecting how closely the comment resembles authentic human expression. These two dimensions directly align with the objective of measuring user-like expressive quality and have shown strong discriminative power across different models in our experiments.

Blind and Independent Annotation Protocol. To ensure fairness, reduce bias, and improve consistency, we adopt a blind, multi-annotator evaluation protocol. For each video, real user comments and model-generated comments are mixed into a unified pool and presented in randomized order. Annotators are only shown the video content and a single comment at a time; the source of the comment (human or model) and the identity of the generating model are fully hidden. Importantly, model-generated comments are *not* directly compared against specific real comments, as such comparison could implicitly reveal comment origin and conflict with the design of the origin prediction task.

Each comment is independently evaluated by multiple annotators using a unified web-based interface (see Figure 5 for an example of the annotation interface). Annotators receive brief instructions and illustrative examples prior to the formal evaluation to ensure a consistent understanding of the criteria. A fixed scoring rubric with clear definitions is used to reduce ambiguity in interpreting human-likeness.

Bias Mitigation and Consistency. Several strategies are employed to mitigate potential annotation bias and variability: (i) multiple independent annotators are used to reduce the influence of individual subjectivity; (ii) randomized presentation is applied to avoid ordering effects; and (iii) disagreement cases are resolved using simple aggregation rules, such as majority voting, with rare highly disputed items reviewed by a more experienced annotator. These practices follow well-established methodologies in open-ended generation benchmarks and improve the stability and reproducibility of the evaluation.

Scalability Considerations. We acknowledge that human evaluation introduces scalability challenges, which are shared by all benchmarks involving open-ended language generation. Due to the diversity and subjectivity of social media comments, it is difficult to define a single ground truth for automatic comparison. Nevertheless, MIMIC-Bench and MIMIC-Data contain multiple high-quality human comments for each video, forming a rich distribution of authentic user responses.

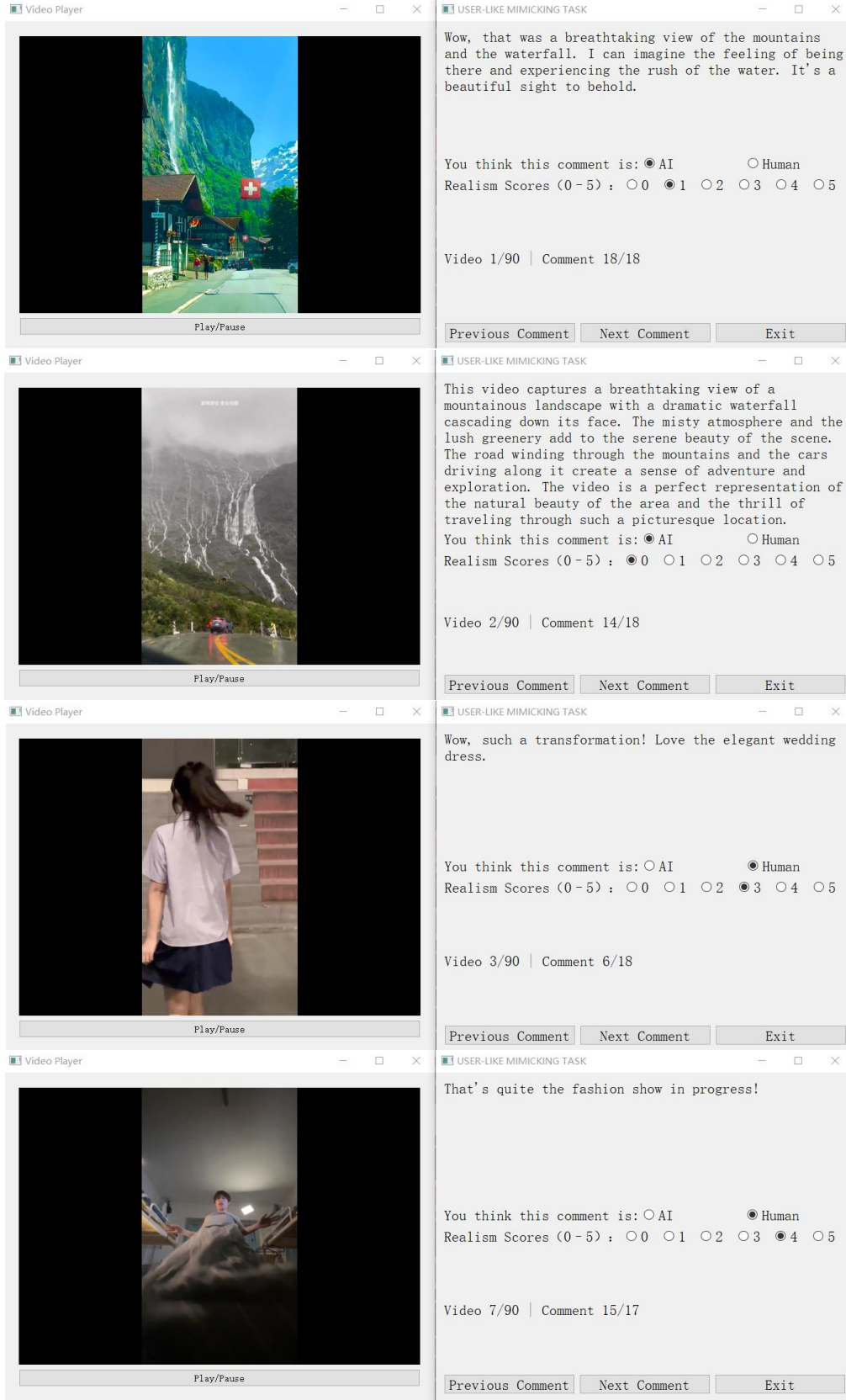


Figure 5: Illustration of the annotation interface used in our **User-like Mimicking Task**. For each video, human annotators are presented with: (1) a displayed video clip, (2) one candidate comment generated either by an MLLM or a human, (3) a forced-choice judgment (*AI* vs. *Human*) on the comment’s origin, and (4) a realism score from 0–5 assessing how human-like the comment appears. We ensure consistent evaluation across all samples. This setup mirrors the model evaluation pipeline in our benchmark and provides a controlled environment for assessing human-centered cognitive and mimicking abilities.

Table 3: Out-of-domain (OOD) evaluation results on videos with different engagement profiles. Performance patterns remain consistent with those observed on MIMIC-Bench.

Model / Tasks	CIU		UIU		CAM			Overall
	TiS	DeS	TaM	ToM	CaM	CoM	CoP	
Qwen2.5-VL-7B	81.1	53.6	79.3	89.6	43.8	58.5	29.3	62.2
InternVL2.5-8B	84.3	51.2	86.8	90.6	49.5	64.5	31.3	65.5
MIMIC-Chat (Ours)	92.5	87.5	87.4	93.2	56.2	78.4	43.8	77.0

This structure provides a foundation for exploring more scalable evaluation strategies in the future, such as hybrid LLM-assisted filtering or distribution-aware metrics, which we identify as an important direction for future work.

E ANALYSIS ON ENGAGEMENT BIAS AND OUT-OF-DOMAIN GENERALIZATION

E.1 RATIONALE OF USING HIGHER-ENGAGEMENT VIDEOS

The inclusion of a subset of higher-engagement videos (top 2–5%) in MIMIC-Bench is motivated by methodological considerations rather than by any specific social or cultural preference. Our primary objective is to ensure content quality, semantic clarity, and the availability of reliable real-user interaction signals, which are essential for evaluating user-like cognition and communicative behavior.

For tasks centered on user comments and human-centric understanding, higher-engagement videos typically exhibit a more stable semantic focus and more representative comment distributions. Such properties are crucial for constructing dependable mimicking and interpretation tasks, as they reduce noise caused by ambiguous content, sparse feedback, or incidental viewer reactions. In contrast, low-engagement videos often lack sufficient or diverse user responses, making it difficult to reliably assess human-aligned reasoning or language behavior.

To mitigate potential biases arising from popularity-based selection, MIMIC-Bench is further curated through a careful multi-stage manual filtering process. This process explicitly avoids stylistic homogeneity and cultural narrowness. The final benchmark spans a wide range of themes, narrative forms, and social contexts, including daily-life moments, instructional content, emotional self-expression, pets, food, travel, outdoor activities, and creative works. As a result, no single style, genre, or cultural pattern dominates the benchmark, even though some videos exhibit higher engagement.

Overall, the use of higher-engagement videos serves as a quality-control mechanism to support reliable human-centric evaluation, rather than as a proxy for social preference or cultural bias. In the following sections, we further examine this choice through out-of-domain and cross-benchmark analyses.

E.2 OUT-OF-DOMAIN EVALUATION ON DIFFERENT ENGAGEMENT PROFILES

To further assess whether the engagement-based video selection introduces unintended bias, we conduct an out-of-domain (OOD) evaluation on videos with engagement profiles different from those used in MIMIC-Bench. Specifically, we construct a small OOD subset by randomly sampling videos with varied and generally lower engagement levels, while preserving the same task formulation and evaluation protocol as the original benchmark.

We evaluate several representative MLLMs, including Qwen2.5-VL-7B, InternVL2.5-8B, and our MIMIC-Chat, on this OOD subset. The results are summarized in Table 3. Although absolute performance varies across models, all three exhibit performance trends that are highly consistent with those observed on MIMIC-Bench.

Table 4: Cross-benchmark results on selected MVBench subtasks.

Model	VideoLLaMA	VideoChatGPT	LLaVA	VideoChat2	InternVL2.5-8B	MIMIC-Chat
Action Antonym	51.0	62.0	63.0	83.5	87.9	90.1
Action Count	34.0	30.5	34.0	39.0	54.4	58.5
Egocentric Navigation	30.0	29.5	27.0	35.0	30.1	35.5

Table 5: Cross-benchmark results on selected TemporalBench subtasks.

Task	InternVL2.5-8B	MIMIC-Chat
Motion Direction / Orientation	58.6	60.4
Action Effector	51.4	52.3
Motion Magnitude	59.5	61.4

Notably, MIMIC-Chat continues to achieve consistent improvements over both Qwen2.5-VL-7B and InternVL2.5-8B across all task groups, including Creator Intent Understanding (CIU), User Interaction Understanding (UIU), and Content Attribute Matching (CAM). The preservation of clear performance margins under distribution shift suggests that the benchmark tasks capture intrinsic model capabilities rather than exploiting specific engagement patterns or video popularity cues.

Overall, the strong agreement between in-domain and OOD evaluations indicates that MIMIC-Bench is not overly sensitive to engagement-based selection, and that MIMIC-Chat’s gains generalize reliably to videos with different interaction statistics.

E.3 CROSS-BENCHMARK GENERALIZATION ANALYSIS

To further evaluate whether the improvements achieved by MIMIC-Chat generalize beyond MIMIC-Bench, we conduct additional experiments on several widely used video reasoning benchmarks, including MVBench and TemporalBench. Although these benchmarks were not explicitly designed to assess human-centric cognition or comment mimicking, some of their subtasks overlap with the cognitive dimensions emphasized in MIMIC-Bench, such as action semantics, temporal reasoning, and egocentric understanding. Therefore, they provide meaningful complementary evidence for cross-benchmark generalization.

Table 4 reports results on representative MVBench subtasks, including Action Antonym, Action Count, and Egocentric Navigation. These tasks involve higher-level human action reasoning, quantity-sensitive inference, and first-person spatial understanding. Across all three subtasks, MIMIC-Chat consistently outperforms multiple baseline models, including its base model InternVL2.5-8B. The observed gains suggest that the improvements induced by MIMIC-Data are not limited to the benchmark setting, but extend to more diverse action- and perspective-oriented reasoning scenarios.

Table 5 presents results on selected TemporalBench subtasks, including Motion Direction/Orientation, Action Effector, and Motion Magnitude. These tasks require fine-grained motion interpretation, recognition of acting body parts, and estimation of physical magnitude. Such skills closely correspond to the content-attribute understanding, human-interaction reasoning, and dynamic semantic extraction emphasized in MIMIC-Bench. MIMIC-Chat again achieves consistent improvements over InternVL2.5-8B across all three subtasks, indicating enhanced temporal and motion reasoning capability.

Overall, despite substantial differences in task formulation, supervision signals, and data distributions between these external benchmarks and MIMIC-Bench, MIMIC-Chat demonstrates stable performance gains across heterogeneous tasks. This cross-benchmark consistency supports the claim that the observed improvements arise from enhanced general reasoning and human-centric abstraction abilities, rather than from overfitting to the structure or data characteristics of MIMIC-Bench.

F DATASET RELEASE, COPYRIGHT COMPLIANCE, AND RISK MITIGATION

This section provides a detailed description of the procedures and safeguards adopted in the construction and release of MIMIC-Data and MIMIC-Bench. In response to the ethical requirements specified in the meta-review, we document our practices for respecting copyright and platform terms of service, preventing the release of identifiable personal information, and mitigating potential risks of reputational harm.

The following subsections describe (1) data source compliance and platform policies, (2) copyright respect and redistribution constraints, (3) privacy protection and de-identification procedures, and (4) safeguards against reputational misuse. These measures are designed to ensure that the benchmark can support academic research while minimizing legal, ethical, and social risks.

F.1 DATA SOURCE AND PLATFORM COMPLIANCE

All videos included in MIMIC-Data were collected exclusively from publicly accessible content on major online video-sharing platforms at the time of collection. We only consider content that is openly viewable without authentication beyond standard platform access and do not access private accounts, restricted materials, paywalled content, or content requiring elevated permissions.

Data collection procedures were designed to comply with the publicly stated terms of service of the respective hosting platforms. We do not employ methods that bypass access controls, circumvent technical safeguards, or scrape non-public data. The dataset construction process focuses on retrieving publicly available video identifiers and associated metadata necessary for benchmark construction and evaluation.

Importantly, MIMIC-Bench does not redistribute raw video files. Instead, we release benchmark annotations, task splits, evaluation prompts, and metadata required for academic research. Users of the benchmark are expected to access original video content directly through the respective hosting platforms under their applicable terms of service.

We also maintain a contact mechanism for content-related concerns. If a content owner or platform requests removal of specific entries from the benchmark, we will review and, where appropriate, remove the corresponding metadata and annotations from future releases.

F.2 COPYRIGHT RESPECT AND REDISTRIBUTION POLICY

MIMIC-Bench and MIMIC-Data do not claim ownership of the original video content hosted on third-party platforms. All copyrights remain with the respective content creators and hosting services. The benchmark is constructed solely for academic research purposes and does not alter or redistribute copyrighted media.

As part of our release policy, we do not provide raw video files, downloadable media content, or tools that facilitate bypassing platform restrictions. Instead, the released materials consist of annotations, benchmark task definitions, evaluation prompts, and structured metadata necessary for reproducible research. Researchers accessing the benchmark are responsible for complying with the terms of service and copyright policies of the original hosting platforms when retrieving video content.

The dataset is released under a research-only usage policy. It is not intended for commercial redistribution, large-scale content replication, or deployment in systems that violate platform policies or infringe upon creator rights. Any downstream use must respect applicable intellectual property regulations and platform-specific rules.

If any copyright holder or platform operator identifies content within the benchmark that raises concerns, we commit to reviewing the request and removing associated metadata or annotations from subsequent releases where appropriate.

F.3 PRIVACY PROTECTION AND DE-IDENTIFICATION PROCEDURES

We adopt multiple layers of safeguards to prevent the inclusion or release of personally identifiable information (PII) in MIMIC-Data and MIMIC-Bench.

First, we do not release uploader identities, account names, profile links, user IDs, contact information, or other metadata that could directly identify individuals. Only benchmark-relevant annotations, task prompts, and non-identifying metadata are retained for research evaluation purposes.

Second, we implement filtering criteria to exclude content that centers on private individuals, minors, or sensitive personal circumstances. Videos that prominently expose private residential addresses, personal contact details, or vulnerable individuals are excluded during dataset construction. We also avoid including content that primarily documents personal disputes, medical conditions, or other highly sensitive life events involving identifiable individuals.

Finally, the released benchmark does not provide structured indexing or retrieval mechanisms designed to profile or track specific individuals. The dataset is organized around task-level evaluation rather than person-level aggregation, thereby reducing the risk of individual targeting.

These measures collectively aim to ensure that the released data does not contain identifiable personal information and does not facilitate individual identification.

F.4 REPUTATIONAL HARM RISK MITIGATION AND RESPONSIBLE USE

In addition to copyright and privacy safeguards, we implement specific measures to reduce the risk that the benchmark could be used in ways that cause reputational harm to individuals or communities.

During dataset construction, we exclude content that contains defamatory statements, targeted harassment, unverified allegations against identifiable persons, or highly sensitive personal topics (e.g., criminal accusations, medical disclosures, or politically charged attacks directed at specific individuals). Videos or associated annotations that center on exposing, shaming, or negatively profiling private individuals are removed based on predefined exclusion criteria.

Furthermore, MIMIC-Bench is not structured to support person-level profiling or tracking. The benchmark is organized around task-based evaluation of multimodal reasoning and language generation, rather than aggregation of content linked to specific creators or individuals. We do not provide tools, indices, or annotations that facilitate the identification or systematic analysis of particular persons.

The dataset is released strictly for academic research and benchmarking purposes. It is not intended for surveillance, behavioral profiling, automated content moderation targeting specific individuals, or any application that may adversely impact the rights, dignity, or reputation of individuals. Users of the benchmark are expected to adhere to applicable legal and ethical standards in downstream use.

We recognize that no large-scale dataset can entirely eliminate misuse risks. Therefore, we commit to monitoring community feedback and revising the benchmark if credible concerns about harmful impact arise. Where appropriate, we will update filtering rules or remove problematic entries in future releases to maintain responsible stewardship of the dataset.