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

Anonymous authors
Paper under double-blind review

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.

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**
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

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, 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
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**

- **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:

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

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 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

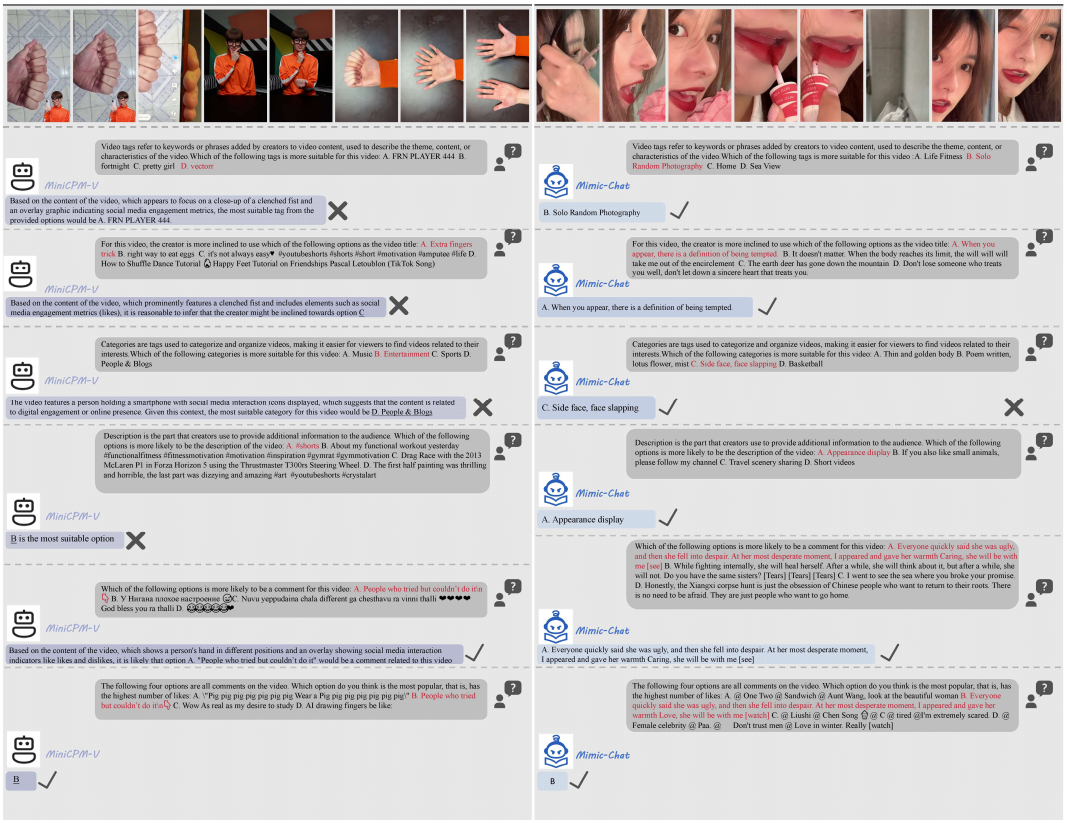

Figure 1: Presentation of the responses to the **MiniCPM-V** and **MIMIC-Chat (Ours)** part of the task. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.

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.

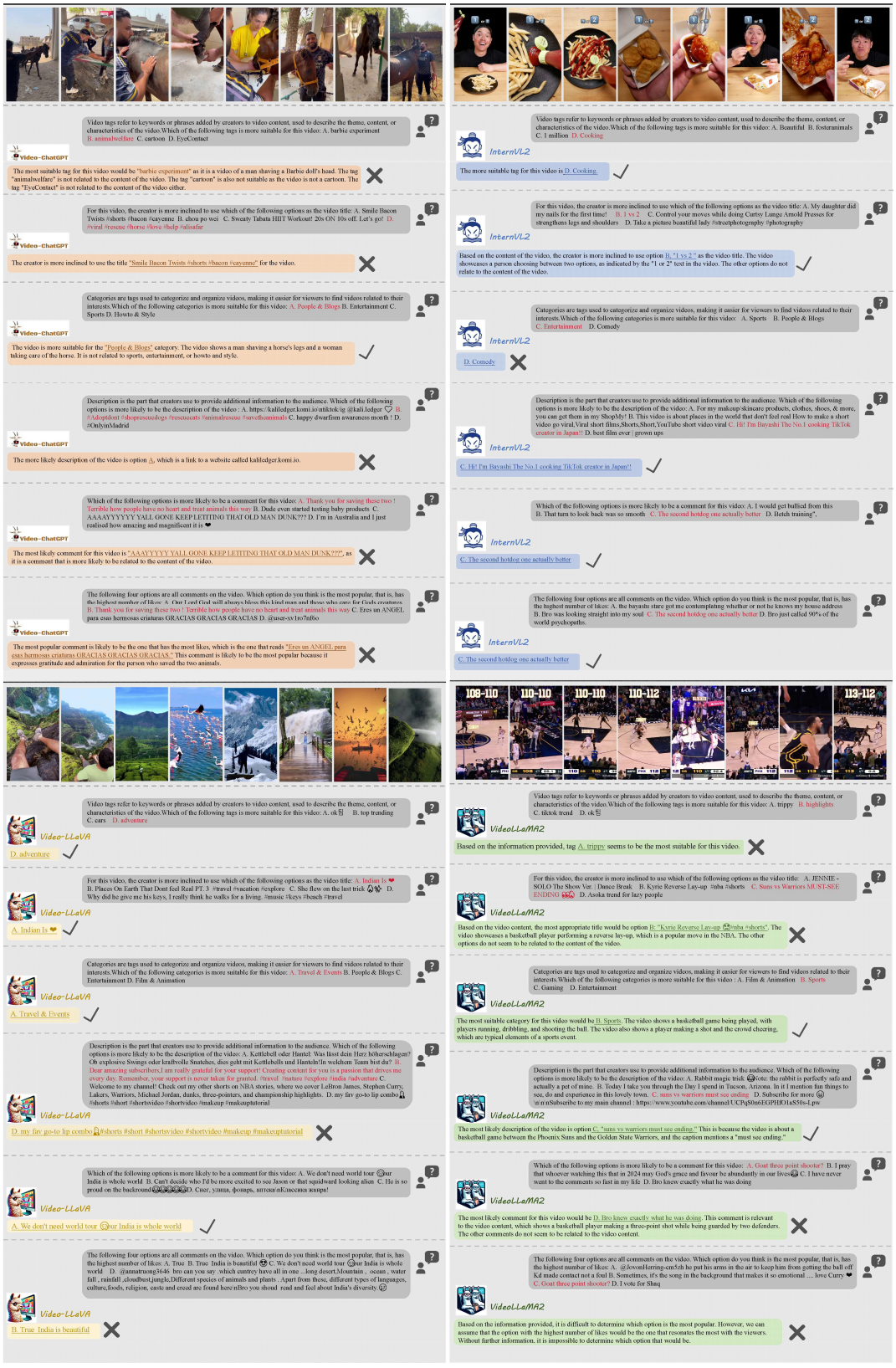

Figure 2: Presentation of the responses to the **Video-ChatGPT**, **InternVL2**, **Video-LLaVA**, and **VideoLLaMA2** part of the task. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.

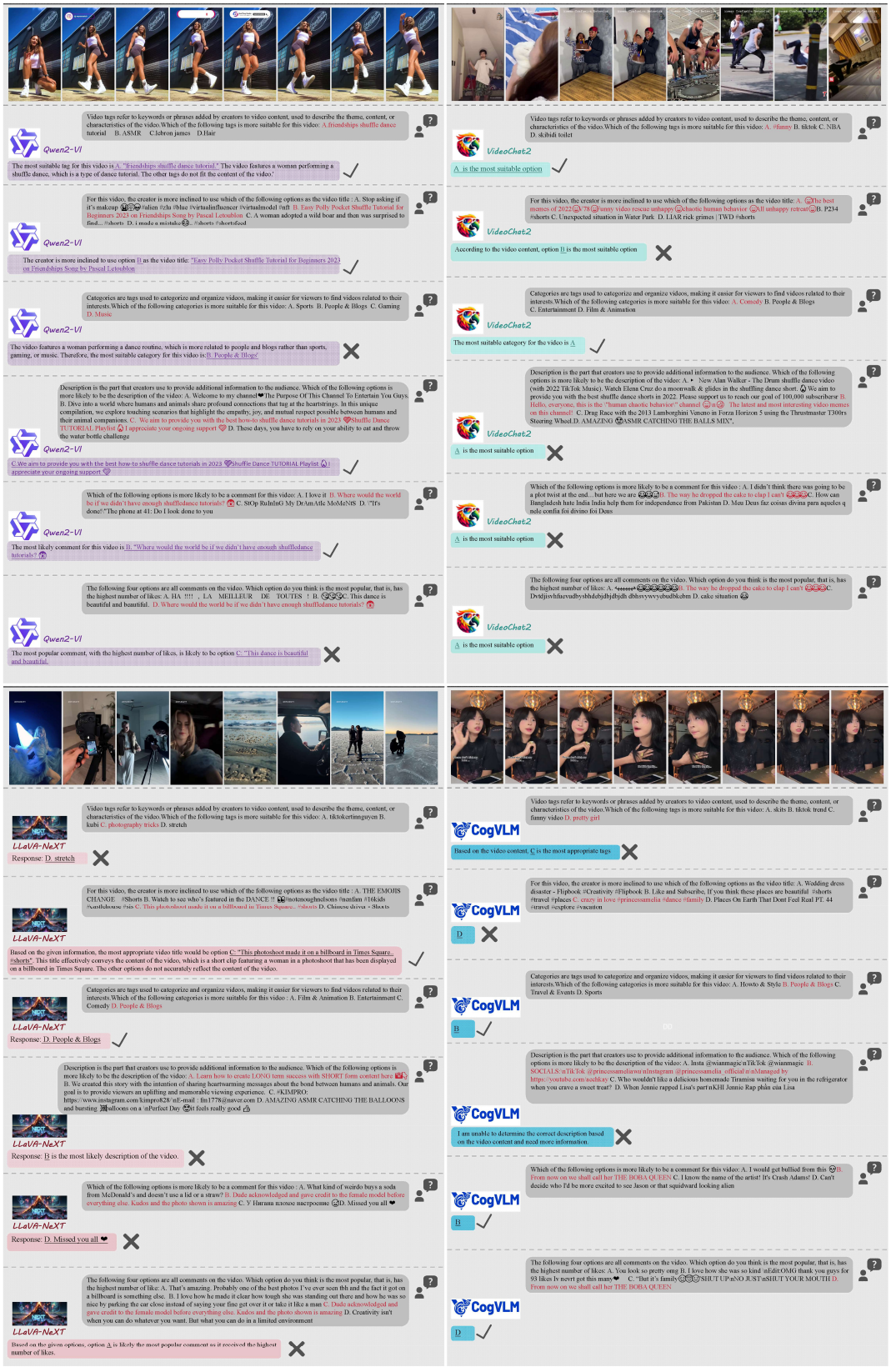

Figure 3: Presentation of the responses to the **Qwen2-VL**, **VideoChat2**, **LLaVA-NeXT**, and **CogVLM2** part of the task. Ground truth is marked in red in the question, and model responses and correctness follow the corresponding question.