FOCUSOMNI: COARSE-TO-FINE TEMPORAL EVIDENCE GROUNDING FOR OMNI-MODAL VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (VLMs) have made significant progress in video reasoning by integrating visual, audio, and textual modalities. However, when long videos contain substantial irrelevant content, existing models that directly reason over uniformly sampled frames often suffer from temporal distraction, as they lack an explicit mechanism to identify which temporal evidence should participate in reasoning. To address this limitation, we propose FocusOmni, a coarse-to-fine temporal evidence grounding framework that enables omni-modal models to identify question-relevant temporal regions before performing multi-modal reasoning, which follows a coarse-to-fine grounding strategy. It first constructs a compact global temporal overview by arranging sampled frames into a grid representation, from which the model identifies the temporal region that contains question-relevant evidence. Based on the grounded temporal region, FocusOmni concentrates visual and auditory evidence within the relevant interval, enabling more effective omni-modal reasoning through explicit temporal evidence grounding. We further introduce ConcatBench, a benchmark that constructs temporally distracted videos by combining question-relevant clips with irrelevant distractor clips, enabling evaluation of segment-level temporal evidence grounding under controlled interference. Extensive experiments on ConcatBench and additional video benchmarks demonstrate that FocusOmni improves both temporal grounding and downstream reasoning performance while requiring only a compact global overview and localized evidence frames. Our code is available at `https://anonymous.4open.science/r/FocusOmni-53C5`.

1 INTRODUCTION

Recent advances in large vision-language models, or VLMs, have enabled substantial progress in video understanding. Models such as Qwen2.5-Omni (Xu et al., 2025) and HumanOmniV2 (Yang et al., 2025) integrate visual, audio, and textual modalities within a unified architecture and generate structured reasoning chains in natural language. HumanOmniV2 further introduces a three-tag output format that organizes description, step-by-step reasoning, and the final answer into a traceable chain of thought.

Despite this progress, existing VLMs still typically process video frames uniformly by sampling a fixed number of frames at equal intervals over the entire video. For long videos containing content unrelated to the question, this strategy forces the model to distribute its attention across irrelevant segments. Consequently, both the visual frames and the audio track may contain irrelevant content, causing the reasoning chain to rely on distracting evidence and potentially producing answers that correspond to irrelevant clips. This limitation has motivated query-aware frame selection and temporal search methods for long-video understanding (Zhang et al., 2025; Ye et al., 2025).

A central obstacle is the lack of an explicit temporal grounding interface. These models describe what they see, reason about the observations, and produce an answer, but they do not explicitly identify where the supporting evidence occurs in the video. The performance degradation stems from the combination of distractors and a fixed sampling schedule that must cover the entire timeline. An explicit grounding interface provides a direct mechanism for addressing this limitation. Once the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

model identifies the relevant temporal interval, subsequent inference can allocate visual computation to that region, allowing the supporting evidence to be observed at a higher temporal density.

We address this gap with FocusOmni, a coarse to fine temporal evidence grounding framework that enables omni modal models to identify question relevant temporal regions before performing multimodal reasoning. First, we introduce a four tag structured output format that provides the model with explicit temporal grounding capability. We extend the three-tag format with a temporal grounding prefix that specifies the question-relevant temporal window in seconds. The model learns this format through LoRA-based supervised fine-tuning on the training split of ConcatBench, enabling it to identify the relevant temporal region before performing detailed video understanding.

To construct a compact global temporal overview in long videos, we project the temporal timeline onto a two dimensional grid image, avoiding the need to reason directly over the raw temporal sequence. Multimodal models are generally more effective at localizing objects within spatial layouts than at indexing moments along a temporal timeline. Arranging sampled frames in a grid therefore allows the model to perceive both the visual content and the spatial position of each frame, while overlaid time-range labels map each spatial position back to its corresponding temporal interval.

Based on the grounded temporal region, FocusOmni concentrates visual and auditory evidence within the relevant interval for fine grained omni modal reasoning. During inference, the model first examines the grid image to predict a coarse segment in Stage 1, then samples dense frames within the predicted interval using a frame rate determined by the segment duration, concentrating the fixed frame budget on the relevant region in Stage 2. This strategy allocates visual tokens according to temporal relevance rather than a fixed frame rate, improving the understanding of long and concatenated videos under a limited computational budget. To preserve the omni modal reasoning capability of the base model, we also truncate the audio track to the predicted segment, keeping visual and audio inputs temporally aligned during Stage 2. Audio is not used for localization because Stage 1 is vision only, but is retained for Stage 2 reasoning over synchronized evidence from the same temporal interval. Together, these components form a unified framework that first locates relevant evidence and then performs detailed multimodal reasoning within the identified region.

Our contributions are summarized below:

- We introduce FocusOmni, a coarse-to-fine temporal evidence grounding framework that enables omni-modal video models to identify question-relevant temporal regions before multimodal reasoning. By explicitly exposing temporal evidence as an intermediate representation, FocusOmni bridges temporal grounding and downstream video reasoning.
- We develop a two-stage grounding mechanism that first builds a compact global temporal overview for coarse evidence localization and then focuses multimodal processing on the grounded region through localized visual and auditory evidence acquisition.
- We construct ConcatBench, a benchmark for evaluating temporal evidence grounding under controlled temporal distraction. By concatenating question-relevant clips with distractor clips and providing automatic segment-level annotations, ConcatBench enables systematic evaluation of a model's ability to isolate relevant temporal evidence.
- Extensive experiments on ConcatBench and additional video benchmarks demonstrate that FocusOmni consistently improves both temporal evidence grounding and downstream answer accuracy, showing the effectiveness of explicit temporal evidence modeling for robust omni-modal video reasoning.

2 Related Work

Large Vision Language Models for Video Understanding. The development of VLMs has progressed from image-only models such as LLaVA (Liu et al., 2023) and InstructBLIP (Dai et al., 2023) to video-capable architectures. VideoChat (Li et al., 2023) and Video ChatGPT (Maaz et al., 2024) extend image-based LLMs with video encoders, while Video LLaMA (Zhang et al., 2023) introduces audio-visual instruction tuning. More recent unified architectures integrate multiple modalities within a single transformer backbone. Qwen2.5-Omni (Xu et al., 2025) jointly processes visual, audio, and textual modalities, while HumanOmniV2 (Yang et al., 2025) builds on this framework with a three-tag reasoning format and multi-stage SFT and GRPO. Our work extends HumanOm-

niV2 with a fourth segment tag for temporal grounding while preserving its omni-modal reasoning capability.

Frame Sampling for Efficient Video Processing. Uniform sampling remains common in video LLMs but can miss query-related temporal cues in long videos (Zhang et al., 2025; Ye et al., 2025). Adaptive methods instead allocate computation to more informative evidence. AdaFrame (Wu et al., 2019) learns an input-dependent frame selection policy, while Q-Frame (Zhang et al., 2025) uses the input query to guide frame selection. MovieChat (Song et al., 2024) compresses dense visual tokens into sparse memories to preserve long-range context. In contrast, our strategy uses the model's predicted segment to determine where to sample and applies the same temporal interval to both dense visual sampling and audio cropping.

Locate then Reason and Temporal to Spatial Projection. SeViLA (Yu et al., 2023) selects keyframes with a localizer before reasoning, while TimeChat (Ren et al., 2024) uses timestamp-aware modeling to ground events temporally. The T* framework Ye et al. (2025) performs temporal search by organizing sampled frames into grids and using spatial search to iteratively update temporal sampling. Similarly, FocusOmni employs grid-based temporal search, but predicts a continuous evidence window $[\hat{t}_s, \hat{t}_e]$ and uses it to concentrate dense visual sampling and crop synchronized audio for Stage 2 omni-modal reasoning.

Audio Visual Temporal Alignment. Audio-visual temporal alignment is central to audio-visual event localization (Tian et al., 2018) and time-aware omni-modal understanding of long videos (Geng et al., 2025). In long or concatenated videos, relevant content may occupy only a narrow temporal window, while the full-length audio track can introduce irrelevant information. Our segment-guided audio truncation extracts audio from the predicted temporal segment $[\hat{t}_s, \hat{t}_e]$ and keeps the audio and visual modalities temporally synchronized during Stage 2 reasoning.

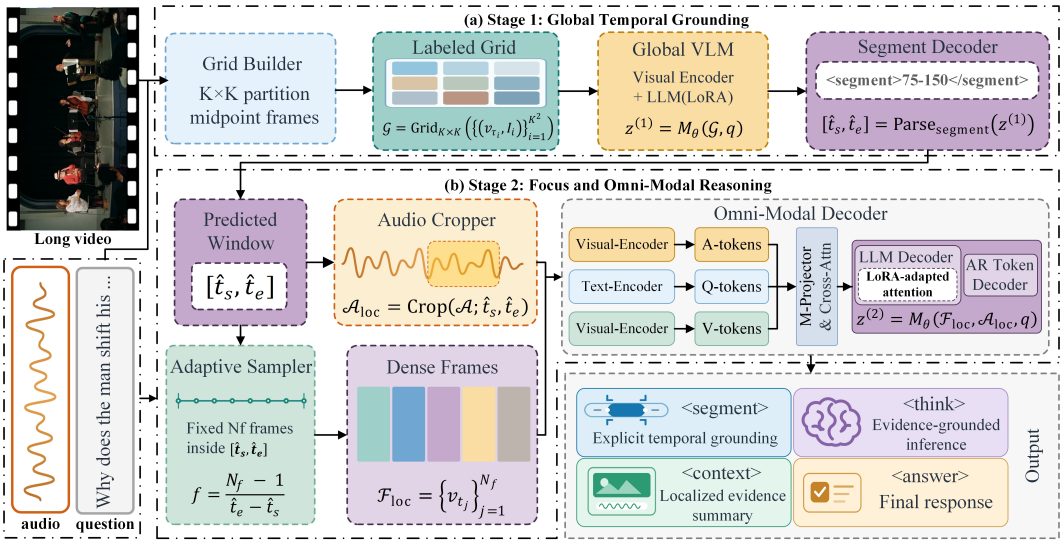
3 METHOD

To address the inefficiency and missing evidence caused by uniform sampling, FocusOmni adopts a coarse to fine temporal evidence grounding strategy. It first constructs a compact global temporal overview to identify where relevant evidence is likely to occur and then focuses its computational resources on that grounded region for local multimodal reasoning. Formally, given a video of duration T, we denote its synchronized visual and audio streams by $\mathcal{V} = \{v_t\}_{t=0}^{T}$ and $\mathcal{A} = \{a_t\}_{t=0}^{T}$, respectively, together with a natural-language question q. The objective is to identify a critical temporal segment $[\hat{t}_s, \hat{t}_e] \subseteq [0, T]$ that contains the evidence needed to answer q, and to reason only over the visual and auditory evidence aligned with this interval. FocusOmni performs this task in two stages. Stage 1 projects the temporal stream into a spatial grid and performs temporal grounding, while Stage 2 reallocates a fixed visual budget to the predicted window and synchronizes the corresponding audio for local omni-modal reasoning.

Figure 1 visualizes the complete pipeline. The predicted window $[\hat{t}_s, \hat{t}_e]$ serves as the explicit bridge between the global temporal overview and fine grained local reasoning. It converts the global search result into a localized evidence region, allowing the model to allocate its visual and auditory processing budget to the portion of the video most likely to contain the answer. The following subsections describe the temporal grounding and adaptive sampling components of this locate-then-focus process in detail.

3.1 FOUR-TAG STRUCTURED OUTPUT

While HumanOmniV2 produces three-tag responses containing context, think, and answer fields, FocusOmni adds an explicit temporal segment prefix before the reasoning chain. The segment prediction specifies the question-relevant window as s_{start}-s_{end} in seconds with one-decimal precision, for example 70.5-130.5. The context section summarizes the visual and auditory content within the localized clip. The think section contains the step-by-step CoT reasoning process based on multimodal cues. The answer section provides the predicted option or final answer. Equivalently, we write the supervision target as $y = \langle s, c, r, a \rangle$ for the segment, context, reasoning, and answer fields. The segment field is generated first and serves as the signal used to construct the local evidence


(a) Stage 1: Global Temporal Grounding
Grid Builder
K×K partition
midpoint frames
Labeled Grid
Global VLM
Visual Encoder
+ LLM(LoRA)
Segment Decoder
75-150
Long video
(b) Stage 2: Focus and Omni-Modal Reasoning
Predicted
Window
Audio Cropper
Omni-Modal Decoder
Visual-Encoder
A-tokens
Text-Encoder
Q-tokens
Visual-Encoder
V-tokens
M-Projector
& Cross-Attn
LLM Decoder
LoRA-adapted
attention
AR Token
Decoder
Adaptive Sampler
Fixed Nf frames
inside
Dense Frames
audio
question
Why does the man shift his ...

Explicit temporal grounding
<think>
Evidence-grounded
inference
<context>
Localized evidence
summary
<answer>
Final response
Output

Figure 1: Overview of FocusOmni. Given a long video and a question, Stage 1 converts the temporal stream into a labeled $K \times K$ grid and predicts a relevant temporal window $[\hat{t}_s, \hat{t}_e]$. Stage 2 allocates a fixed frame budget within this window, crops the synchronized audio to the same interval, and performs local omni-modal reasoning to generate a structured four-tag response.

set during inference. By decoupling temporal localization from multimodal reasoning, this output format establishes a progressive reasoning pathway while remaining backward compatible with HumanOmniV2. Removing the initial segment prefix restores the original three-tag structure.

3.2 Grid-Based Two-Stage Inference

During inference time, FocusOmni replaces traditional uniform frame sampling with a two-stage pipeline that combines grid-guided spatial localization with adaptive dense sampling. Stage 1 is vision-only and takes the grid image together with the question as input. It projects the temporal timeline into a two-dimensional grid image, avoiding the need to reason directly over the raw temporal sequence. Given a video of duration T, we divide its timeline into $N_g = K^2$ equal-length temporal intervals. The i-th interval is defined as

$$I_i = [(i-1)T/N_g,\ iT/N_g], \qquad i \in \{1, \ldots, N_g\}. \tag{1}$$

For each interval, we extract a representative frame at its temporal midpoint,

$$\tau_i = (i - \tfrac{1}{2})T/K^2, \tag{2}$$

and arrange the resulting frames in row-major order into a labeled $K \times K$ grid given by

$$\mathcal{G} = \mathrm{Grid}_{K\times K}\left(\{(v_{\tau_i}, I_i)\}_{i=1}^{K^2}\right). \tag{3}$$

Each grid cell retains its corresponding time-range label I_i. This projection allows the model to inspect the entire video timeline at once while maintaining a direct mapping between each spatial cell and its corresponding temporal interval. It therefore leverages the stronger sensitivity of multimodal models to spatial layouts than to temporal order.

Conditioned on the grid and question, the first-stage model produces

$$z^{(1)} = M_\theta(\mathcal{G}, q), \tag{4}$$

from which we parse the segment field to obtain the predicted evidence interval, expressed as

$$[\hat{t}_s, \hat{t}_e] = \mathrm{Parse}_{\mathrm{segment}}\left(z^{(1)}\right). \tag{5}$$

This step does not produce the final answer. It identifies a compact window that restricts subsequent visual sampling and audio processing to the likely evidence region.

In Stage 2, FocusOmni performs fine-grained omni-modal reasoning within the predicted segment $[\hat{t}_s, \hat{t}_e]$, where the predicted window must satisfy $0 \le \hat{t}_s < \hat{t}_e \le T$. Let N_f denote the fixed local frame budget. We distribute the N_f samples uniformly across the predicted window, taking its two endpoints as the first and the last sample, so that

$$t_j = \hat{t}_s + \frac{j-1}{N_f - 1}(\hat{t}_e - \hat{t}_s), \qquad j \in \{1, \dots, N_f\}, \tag{6}$$

which corresponds to an effective sampling rate of

$$f = \frac{N_f - 1}{\hat{t}_e - \hat{t}_s}. \tag{7}$$

Because $\hat{t}_e - \hat{t}_s$ varies across samples, f is not fixed. A short predicted window is sampled densely and a long one sparsely, while the number of sampled frames remains N_f. This yields the localized visual set

$$\mathcal{F}_{\text{loc}} = \left\{ v_{t_j} \right\}_{j=1}^{N_f}. \tag{8}$$

This places the full frame budget within the predicted window, increasing observation density for short actions and fine-grained details while avoiding redundant processing of irrelevant portions.

To preserve omni-modal coherence, we crop the synchronized audio stream to the predicted window, written as

$$\mathcal{A}_{\text{loc}} = \text{Crop}(\mathcal{A}; \hat{t}_s, \hat{t}_e). \tag{9}$$

The second-stage model then reasons over the temporally aligned local visual and audio evidence, producing

$$z^{(2)} = M_\theta(\mathcal{F}_{\text{loc}}, \mathcal{A}_{\text{loc}}, q). \tag{10}$$

Because both inputs are crops of the original streams, they no longer carry their position on the global timeline. We therefore pass the predicted window back into the Stage 2 prompt, which states in natural language the time range occupied by the localized evidence in the original video, for example by stating that the provided clip corresponds to seconds $\hat{t}_s$ through $\hat{t}_e$. Stage 2 can thus refer to absolute timestamps while observing only the localized window.

The final structured response is decoded as $\hat{y} = \text{Decode}(z^{(2)})$. Because both modalities originate from $[\hat{t}_s, \hat{t}_e]$, this avoids cross-segment semantic interference from pairing locally sampled frames with the full audio stream. If the Stage 1 segment field cannot be parsed or violates $0 \le \hat{t}_s < \hat{t}_e \le T$, the system falls back to uniform sampling over the whole video with the same frame budget. This forfeits the coarse-to-fine benefit for that sample while the Stage 1 pass has already been paid for, as discussed in Appendix E.

3.3 Training Strategy

FocusOmni is trained with a single stage of LoRA-based supervised fine-tuning (Hu et al., 2021) to learn the four-tag output format. For a training instance x with target sequence $y = \langle s, c, r, a \rangle$, we optimize the standard autoregressive objective over all target tokens, written as

$$\mathcal{L}_{\text{SFT}} = -\sum_{\ell=1}^{L} \log p_{\theta+\Delta\theta}\left(y_\ell \mid y_{<\ell}, x\right). \tag{11}$$

We train on the full sequence to enable the model to learn the complete autoregressive structure and avoid the degenerate generation behavior observed with partial masking strategies.

LoRA parameterizes each adapted projection matrix as

$$\Delta W = BA, \qquad \text{rank}(\Delta W) \le r. \tag{12}$$

We use rank $r = 16$ and scaling factor $\alpha = 32$ for the query, key, value, and output projections W_q, W_k, W_v, W_o of the language model's attention layers, while keeping all other parameters frozen, including the vision encoder, audio tower, feed-forward layers, and token embeddings. The training data $\mathcal{D}_{\text{SFT}}$ is drawn from the training split of ConcatBench described in Section 4, which contains 2,056 video-question pairs. Here, x_i denotes a conversation consisting of a system prompt, video frames, and a user question, while y_i denotes the corresponding gold four-tag solution string.

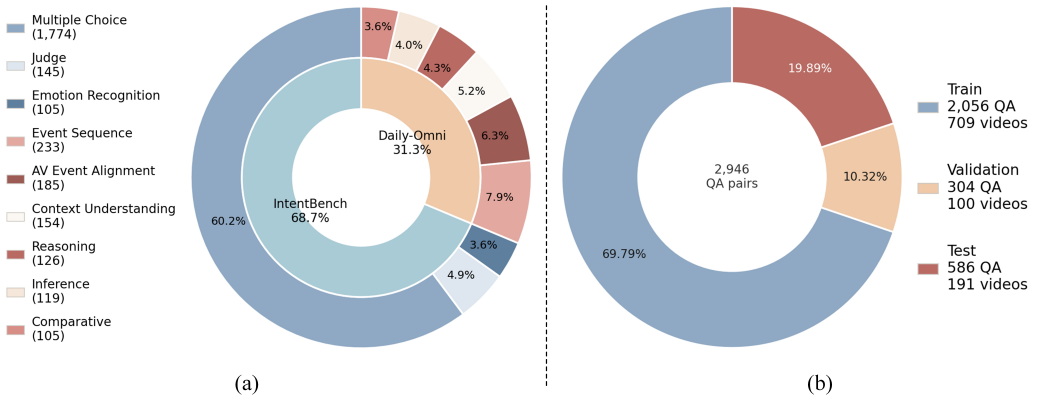


Figure 2: ConcatBench composition and split. (a) Source datasets and problem types of asked clips. (b) Train, validation, and test partition. Video counts refer to concatenated videos, and QA counts refer to question-answer pairs.

Each pair contributes a four-tag training sample and also serves as a candidate for a three-tag replay sample. The replay set written to disk is obtained by uniformly random subsampling from these candid1ates and is not selected based on difficulty or loss. Validation and test videos are never used during SFT. Training uses a learning rate of 1×10^{-5} with cosine decay and 5% linear warmup, a per-device batch size of 1 with 8-step gradient accumulation for an effective batch size of 8, and 2,000 total training steps.

4 The ConcatBench Benchmark

Construction. ConcatBench constructs temporally distracted videos by combining a question relevant clip with one to four irrelevant distractor clips. The position of the relevant clip is uniformly randomized, and the total duration of the resulting video is capped at 150 seconds, preventing models from exploiting positional priors and requiring them to localize the evidence window before answering. The question-relevant clips are drawn from IntentBench and Daily-Omni, two public omni-modal video QA datasets whose original questions and annotations are preserved. The benchmark contains 1,000 concatenated videos and 2,946 QA pairs, and Figure 2(a) summarizes the source datasets and problem types of the asked clips.

Evaluation. ConcatBench evaluates segment level temporal evidence grounding and downstream answer accuracy under controlled temporal interference. The segment annotation marks the boundaries of the question-relevant constituent clip within the concatenated video, so the benchmark evaluates localization of that clip under temporal distraction. It does not evaluate identification of a minimal evidence interval, which may be shorter than the clip containing it. We discuss this distinction in Appendix E. Segment annotations are derived automatically during splicing, allowing the benchmark to scale without human annotation. Appendix B provides complete details on source datasets, splicing, annotation, file format, and statistics.

Splits. Before any fine tuning, we fix the train, validation, and test partition shown in Figure 2(b). QA pairs are grouped by shared source clips so that no source clip asked about in the test split appears in training or validation. Checkpoint selection uses validation, and every result in Section 5.2 is computed on the held out test split described in Appendix B.3.

5 Experiments

5.1 Experimental Setup

We fine tune the HumanOmniV2 model built on Qwen2.5 Omni 7B using LoRA SFT on the training split of ConcatBench. Our main evaluation reports answer accuracy and temporal IoU on Concat-

Bench, together with transfer results on non concatenated omni modal benchmarks. All Concat-Bench results use the held out test split with checkpoint selection on validation, except for the controlled localization intervention in Section 5.3 and the localization quality analysis in Appendix D, which use the full benchmark pool. All baselines in Table 1 use eight uniformly sampled frames and video only input, while at inference FocusOmni uses a 3×3 Stage 1 grid and the same number of adaptively sampled Stage 2 frames with synchronized audio, so part of the performance difference reflects the audio modality. HumanOmniV2's published default of 32 frames is cited only as a reference for visual input and is not used for any result in Table 1. Complete training, baseline, metric, and implementation details are provided in Appendix C.

5.2 Main Results on ConcatBench

Table 1: Comparison with other methods on the held-out test split of ConcatBench. All baselines use eight uniformly sampled frames and video-only input, whereas FocusOmni uses the same number of localized frames together with the synchronized audio segment.

Methods	LLM Size	Multiple Choice	Event Sequence	AV Event Alignment	Context Understanding	Judge	Reasoning	Inference	Emotion Recognition	Comparative	Avg
Proprietary MLLMs											
GLM-4.6V (GLM-V Team et al., 2025)	106B	60.0	25.3	24.3	27.3	60.1	21.4	18.4	8.6	23.8	43.8
Qwen-VL-Max (Bai et al., 2023)	–	54.4	20.6	24.9	23.4	54.4	22.2	24.4	15.2	26.7	40.6
Open-Source MLLMs											
Qwen2-VL (Wang et al., 2024)	2B	50.0	8.5	4.3	1.3	0.0	7.1	2.5	10.5	0.0	31.8
Qwen2.5-VL (Qwen Team et al., 2025)	7B	56.4	6.0	2.7	5.1	0.0	3.1	2.5	14.2	0.9	35.6
VideoLLaMA2 (Cheng et al., 2024)	7B	44.6	23.2	26.5	24.0	46.5	24.0	**25.2**	7.6	19.0	37.0
Ola (Liu et al., 2025)	7B	56.1	**26.6**	**27.0**	27.3	48.9	23.0	16.0	22.9	12.4	44.4
Qwen2.5-Omni (Xu et al., 2025)	7B	56.6	25.3	30.8	24.0	53.9	**32.5**	29.4	26.7	20.0	46.2
HumanOmniV2 (Yang et al., 2025)	7B	65.5	6.9	3.9	1.3	**62.0**	0.0	1.7	33.3	1.0	44.6
FocusOmni (Ours)	7B	**66.6**	24.5	23.8	**29.2**	58.6	25.4	**25.2**	**30.5**	**35.2**	**52.4**

Table 1 presents the main results. FocusOmni achieves an average accuracy of **52.4%**, improving by **+7.8%** over HumanOmniV2 at 44.6% and by +8.9% over Qwen2.5-Omni at 43.5%. The 7B model also exceeds the reported score of GLM-4.6V at 106B with 43.8%. The gains are concentrated on temporally demanding tasks, where FocusOmni reaches 24.5% on Event Sequence, 25.4% on Reasoning, and 35.2% on Comparative, compared with 6.9%, 0.0%, and 1.0% for HumanOmniV2. HumanOmniV2 performs higher only on Judge, at 62.0% versus 58.6%, likely because binary judgments can sometimes be inferred from global statistics without precise localization.

5.3 Ablation Studies

We retain three analyses in the main paper. The grid-resolution ablation examines the key design choice underlying the temporal-to-spatial projection. The component-wise ablation removes one component at a time from the full model to measure the individual contributions of grid-based Stage 1 localization and segment-guided audio. The controlled localization experiment intervenes on the evidence window while keeping the rest of the Stage 2 pipeline fixed. Per-category, localization-quality, and loss-masking analyses are reported in Appendix D.

Grid Size. We evaluate grid resolutions $K \in \{2, 3, 4, 5\}$ against a uniform sampling baseline with 8 frames and no temporal projection. Figure 3(a) reports three complementary metrics. Segment IoU is shown on the left y axis as a blue line, grid image size in megapixels on the right y axis as gray bars, and marginal efficiency on the far right y axis as a green dashed line. Each grid cell is 340×340 px, yielding $(K \cdot 340)^2$ total pixels. We define the marginal efficiency as the per-megapixel IoU gain over the previous grid resolution,

$$\eta_{K_i} = \frac{\text{IoU}(K_i) - \text{IoU}(K_{i-1})}{\text{MP}(K_i) - \text{MP}(K_{i-1})}, \tag{13}$$

where $\text{MP}(K) = (K \cdot 340)^2 / 10^6$, and IoU is expressed in percent, so that η measures the IoU gain per megapixel. This metric isolates whether additional visual tokens yield commensurate localization improvement.

All grid-based variants outperform the uniform baseline by +3.8% to +6.9% in IoU, confirming that temporal-to-spatial projection provides a stronger inductive bias for temporal grounding. Segment IoU rises sharply from 2×2 at 35.71% to 3×3 at 38.44%, then plateaus. The 4×4 configuration

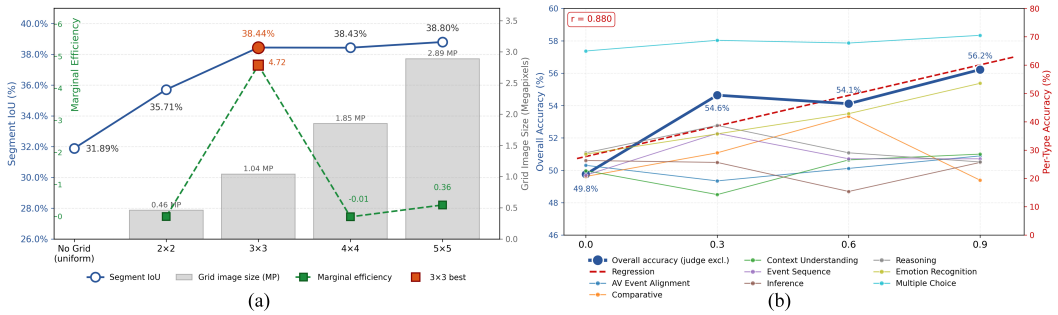


Figure 3: Grid resolution ablation and controlled localization. (a) Segment IoU (blue), grid image size (gray bars), and marginal efficiency (green dashed, Eq. 13) against grid resolution. (b) Accuracy under a forced-IoU segment. The dashed line is the linear reference, and the thin lines show per-problem-type accuracy.

reaches 38.43%, while 5×5 reaches only 38.80%. The marginal efficiency metric makes this pattern explicit, with $\eta_{3\times3} = 4.72$ dominating $\eta_{4\times4} = -0.01$ and $\eta_{5\times5} = 0.36$. Thus, the transition from 2×2 to 3×3 is the only expansion that yields meaningful per-megapixel improvement. Beyond this point, the larger image size, from 1.04 MP at 3×3 to 2.89 MP at 5×5, provides negligible localization gain. We therefore adopt $K = 3$.

Component Wise Ablation. In this ablation, each configuration is derived from the full model by removing exactly one component. Thus, each row differs from the full model by a single factor, allowing the reported deltas to be attributed to that factor. Results are shown in Table 2.

Table 2: Ablation of inference components on the held-out ConcatBench test split.

Configuration	Acc.	IoU
FocusOmni (full, grid + audio)	**52.4**	**38.4**
w/o grid	48.9	31.8
w/o audio	49.5	38.4

Fine-tuning HumanOmniV2 to generate the four-tag response enables segment prediction. The non-tuned backbone produces no segment field and obtains an IoU of 0.0%, whereas the fine-tuned model with uniform Stage 1 sampling already reaches 31.8%. Because this comparison also differs in input modality and training data, we treat it as descriptive rather than a controlled estimate of the accuracy contribution from the four-tag objective. Removing the grid and reverting to uniform Stage 1 sampling costs 3.5% accuracy and 6.6% IoU, confirming that the temporal-to-spatial projection is the component responsible for localization in the full pipeline.

Removing segment-guided audio costs 2.9% accuracy, while segment IoU remains 38.4%. This is expected because audio does not participate in grid-based spatial search. Instead, audio contributes during dense reasoning in Stage 2 by providing evidence that is not visible in individual frames, including speech, prosody, and audio-visual event consistency. Truncating the audio to the predicted segment also removes auditory distractors from unrelated clips in ConcatBench.

Controlled Localization Experiment. To isolate the effect of temporal grounding, we replace the model's segment prediction with a forced segment of the ground truth duration, shifted to achieve IoU $t \in 0, 0.3, 0.6, 0.9$. We apply the standard Stage 2 pipeline while keeping the sampling budget, audio crop, prompt, and decoding fixed. Thus the segment window is the only manipulated variable. This experiment is used only for controlled comparison over the full ConcatBench pool, not for held-out test evaluation. Figure 3(b) shows that accuracy increases from 49.8% at IoU $= 0$ to 56.2% at 0.9, with a 6.5% overall gain, although the trend is not strictly monotonic. On 820 shared questions, the paired bootstrap difference between IoU $= 0.9$ and 0.0 is +6.5% (95% CI: [+3.1, +9.9]). These results show that improved temporal localization benefits answer accuracy under a fixed Stage 2 pipeline.

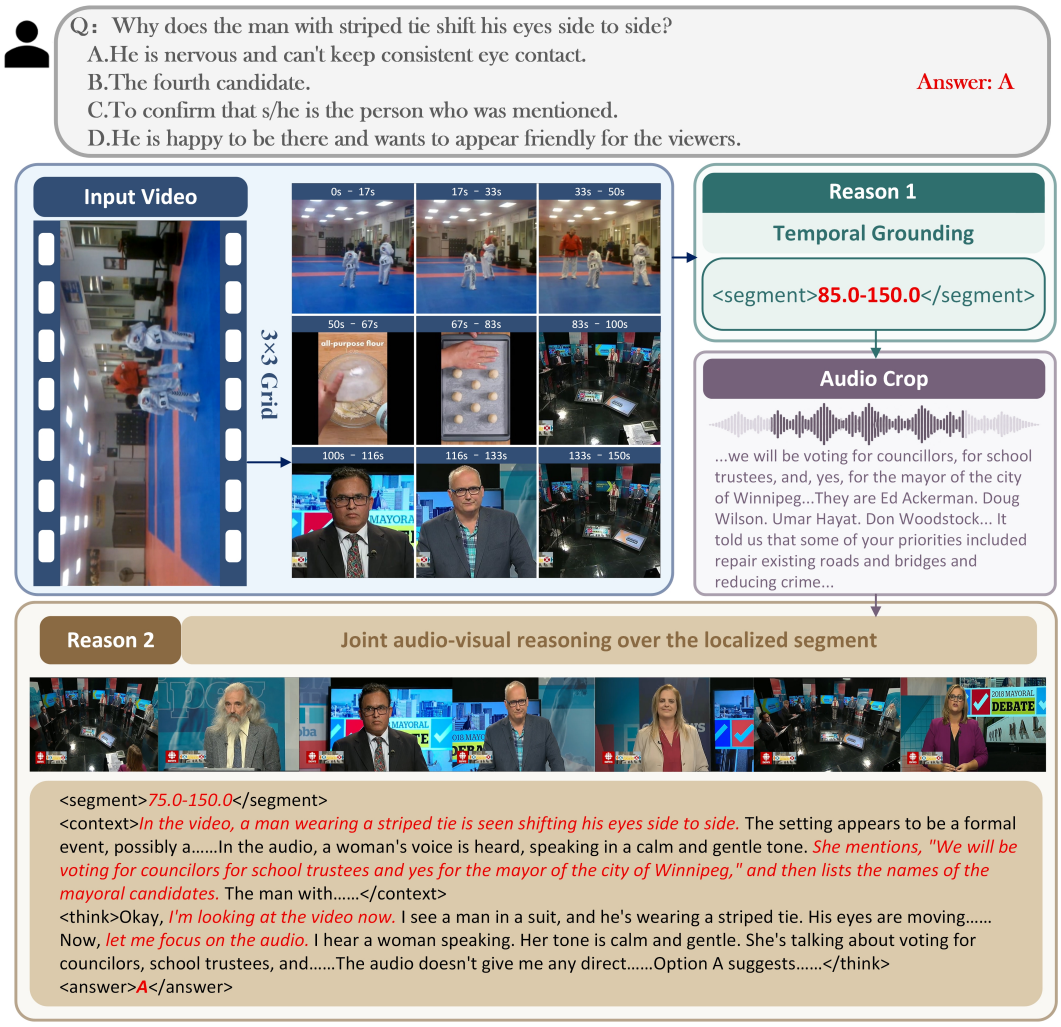

Q: Why does the man with striped tie shift his eyes side to side?
A.He is nervous and can't keep consistent eye contact.
B.The fourth candidate.
C.To confirm that s/he is the person who was mentioned.
D.He is happy to be there and wants to appear friendly for the viewers.
Answer: A
Input Video
3×3 Grid
0s - 17s
17s - 33s
33s - 50s
50s - 67s
67s - 83s
83s - 100s
100s - 116s
116s - 133s
133s - 150s
Reason 1
Temporal Grounding
85.0-150.0
Audio Crop
...we will be voting for councillors, for school trustees, and, yes, for the mayor of the city of Winnipeg...They are Ed Ackerman. Doug Wilson. Umar Hayat. Don Woodstock... It told us that some of your priorities included repair existing roads and bridges and reducing crime...
Reason 2
Joint audio-visual reasoning over the localized segment
75.0-150.0
<context>In the video, a man wearing a striped tie is seen shifting his eyes side to side. The setting appears to be a formal event, possibly a.....In the audio, a woman's voice is heard, speaking in a calm and gentle tone. She mentions, "We will be voting for councilors for school trustees and yes for the mayor of the city of Winnipeg," and then lists the names of the mayoral candidates. The man with.....</context>
<think>Okay, I'm looking at the video now. I see a man in a suit, and he's wearing a striped tie. His eyes are moving..... Now, let me focus on the audio. I hear a woman speaking. Her tone is calm and gentle. She's talking about voting for councilors, school trustees, and.....The audio doesn't give me any direct.....Option A suggests.....</think>
<answer>A</answer>

Figure 4: Visualization of our method on ConcatBench.

Qualitative Example. Figure 4 shows a worked example of the full pipeline on a ConcatBench sample, from the labeled Stage 1 grid and predicted segment to the generated four-tag response. The question asks why the man with a striped tie shifts his eyes from side to side, with option A as the ground-truth answer. Stage 1 lays out the 150-second video as a 3×3 grid with time-range labels and predicts the question-relevant window. Stage 2 samples dense frames within this window and crops the audio to the same interval, enabling reasoning over detailed visual content and synchronized speech. The example illustrates how FocusOmni first grounds the relevant interval and then generates the context, reasoning, and answer fields.

6 Conclusion

We address temporal evidence grounding in long video understanding with FocusOmni, which enables coarse to fine grounding before detailed omni modal reasoning. We first construct a compact global temporal overview with a 2D spatial grid and then concentrate visual and auditory evidence within the grounded temporal region. Lightweight fine-tuning equips the model with explicit segment prediction in its reasoning chain. We will release ConcatBench, a benchmark of concatenated videos with automatically derived segment annotations. On its held-out test split, FocusOmni improves temporal localization and answer accuracy over reproduced baselines while remaining competitive on transfer benchmarks, supporting explicit temporal grounding as a promising direction for omni-modal video reasoning.

REPRODUCIBILITY STATEMENT

The complete method description is given in Section 3, covering the four-tag structured output in Section 3.1, the grid-based two-stage inference procedure together with the adaptive temporal sampling rule and segment-guided audio truncation in Section 3.2, Eqs. 1 through 10, and the LoRA supervised fine-tuning objective and hyperparameters in Eqs. 11 through 12. The construction of ConcatBench is documented in Appendix B, covering source datasets, filler pool, splicing protocol, automatic temporal annotation, file format, full statistics, and the train/validation/test partition. The split is fixed before fine-tuning, the test videos are held out from training, and checkpoint selection uses validation, so every number in Section 5.2 is computed on held-out test data. Training details, baseline configurations, evaluation metrics, and the complete inference configuration of FocusOmni are listed in Appendix C. Additional ablations, localization-quality analysis, and the loss-masking study are reported in Appendix D. The implementation of FocusOmni, the ConcatBench construction pipeline, and all configuration files will be released publicly upon publication.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on experiment design and methodology, to help formalize notation and selected equations in the method section, and to assist in interpreting experimental results, including paired-bootstrap confidence intervals and significance assessments. We did not use generative AI tools to generate synthetic datasets, develop the core conceptual framework of FocusOmni, formulate or prove mathematical claims, clean or reformat datasets, or support qualitative data analysis. Translation and proof-writing tasks are not applicable to this work. We also used generative AI tools to create and edit analysis and plotting code, draft parts of the paper, improve readability, suggest structural and narrative options, and brainstorm. We reviewed all AI-assisted work. AI-generated code was verified and tested by the authors, all numerical results reported in the paper were produced by our experiments and cross-checked against experiment logs, and AI-drafted text was fact-checked before inclusion. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic, 2024.

J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, W. Zhu, W. Zhang, Z. Luo, D. Zhao, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.

G. Comanici, E. Bieber, M. Schaekermann, A. M"uller, A. Anand, P. Barham, S. Borgeaud, S. Bradley, Y. Cui, 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.

W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

C. Fu, H. Lin, X. Wang, Y.-F. Zhang, Y. Shen, X. Liu, H. Cao, Z. Long, H. Gao, K. Li, et al. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Gemini Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

T. Geng, J. Zhang, Q. Wang, T. Wang, J. Duan, and F. Zheng. LongVALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. *arXiv preprint arXiv:2411.19772*, 2025.

GLM-V Team, W. Hong, W. Yu, X. Gu, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

J. Hong, S. Yan, J. Cai, X. Jiang, Y. Hu, and W. Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. *arXiv preprint arXiv:2502.04326*, 2025.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

Z. Lian, H. Sun, L. Sun, H. Gu, Z. Wen, S. Zhang, S. Chen, M. Xu, K. Xu, K. Chen, et al. Explainable multimodal emotion recognition. *arXiv preprint arXiv:2306.15401*, 2023.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Z. Liu, Y. Dong, J. Wang, Z. Liu, W. Hu, J. Lu, and Y. Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025.

J. Lu, C. Clark, S. Lee, Z. Zhang, R. Khosla, R. Marten, D. Hoiem, and A. Kembhavi. Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

M. Maaz, H. Rasheed, S. Khan, and F. S. Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024.

OpenBMB Team. MiniCPM-o: A GPT-4o level MLLM for vision, speech and multimodal live streaming on your phone. *arXiv preprint arXiv:2501.04331*, 2025.

Qwen Team, S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Meng, Z. Zhang, and Z. Wang. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

S. Ren, L. Yao, S. Li, X. Sun, and L. Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, H. Chi, X. Guo, T. Ye, Y. Zhang, Y. Lu, J.-N. Hwang, and G. Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision*, 2018.

P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Z. Wu, C. Xiong, C.-Y. Ma, R. Socher, and L. S. Davis. AdaFrame: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Z. Yang et al. HumanOmniV2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

J. Ye, Z. Wang, H. Sun, K. Chandrasegaran, Z. Durante, C. Eyzaguirre, Y. Bisk, J. C. Niebles, E. Adeli, L. Fei-Fei, J. Wu, and M. Li. T*: Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

S. Yu, J. Cho, P. Yadav, and M. Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, 2023.

A. Zadeh, M. Chan, P. P. Liang, E. Tong, and L.-P. Morency. Social-IQ: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

H. Zhang, X. Li, and L. Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023.

S. Zhang, J. Yang, J. Yin, Z. Luo, and J. Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Z. Zhou, R. Wang, Z. Wu, and Y.-G. Jiang. Daily-Omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

A Qualitative Illustrations

Figure 5 focuses on the input-processing path. Uniform sampling exposes the reasoning module to multiple unrelated clips, whereas FocusOmni first uses a global grid to identify a relevant interval and then extracts dense local frames with aligned audio. The illustration is qualitative and does not imply that every sample follows exactly the same processing path. Figure 6 shows the corresponding reasoning path produced by FocusOmni on a ConcatBench example. The segment field is emitted first, and the context and reasoning fields then refer to the localized clip.

Figure 7 provides a qualitative comparison under temporal distraction. Given the same question and concatenated video, the baseline incorporates evidence from an irrelevant clip and consequently predicts the wrong answer. FocusOmni instead predicts the relevant interval before detailed reasoning, enabling it to focus on the corresponding visual and auditory evidence and answer correctly.

B ConcatBench Construction and Statistics

To rigorously evaluate temporal localization and segment-aware reasoning, we construct ConcatBench. This multi-source concatenated video benchmark draws on the IntentBench (Yang et al., 2025) and Daily-Omni (Zhou et al., 2025) datasets. The original benchmarks contain a single semantic segment per video. ConcatBench instead assembles temporally concatenated video sequences with randomly inserted distractor clips, requiring the model to actively localize the question-relevant interval before answering.

B.1 Dataset Construction

Source Datasets. We leverage two public omni-modal QA datasets as the asked clip sources. IntentBench (Yang et al., 2025) provides 633 videos and 2,689 questions focused on human intention and social interaction. Daily-Omni comprises 684 videos and 1,197 questions from everyday life

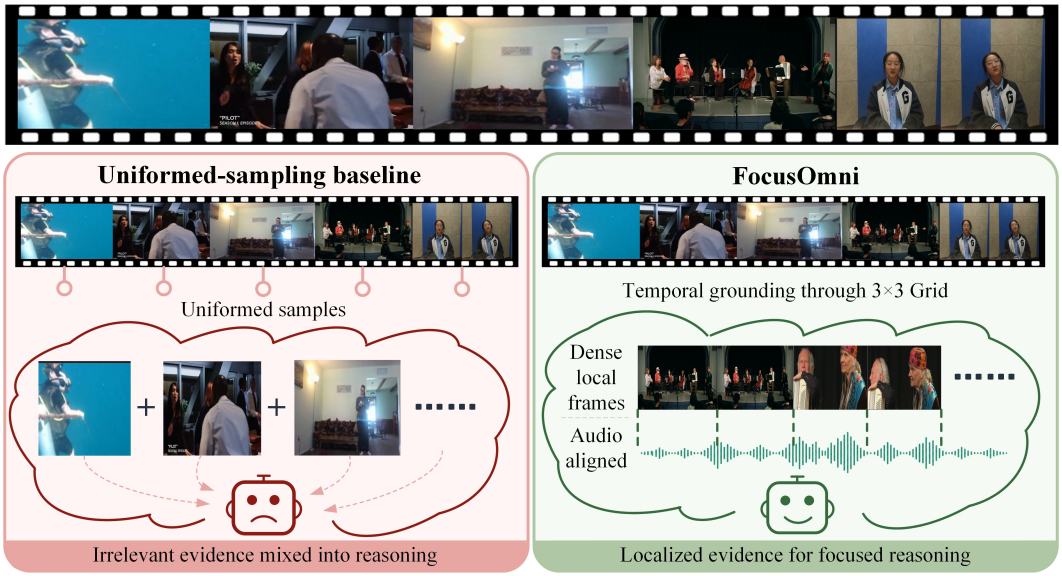


Figure 5: Qualitative comparison of data processing under the same input. The uniform-sampling baseline receives the same concatenated video and question but processes evidence from multiple unrelated segments, which introduces distractors into its reasoning process. In contrast, FocusOmni first performs coarse temporal grounding with a global 3×3 grid, identifies the question-relevant segment, and then extracts dense local frames together with temporally aligned audio. This comparison illustrates the motivation behind FocusOmni. Grounding before detailed multimodal reasoning reduces interference from irrelevant video content.

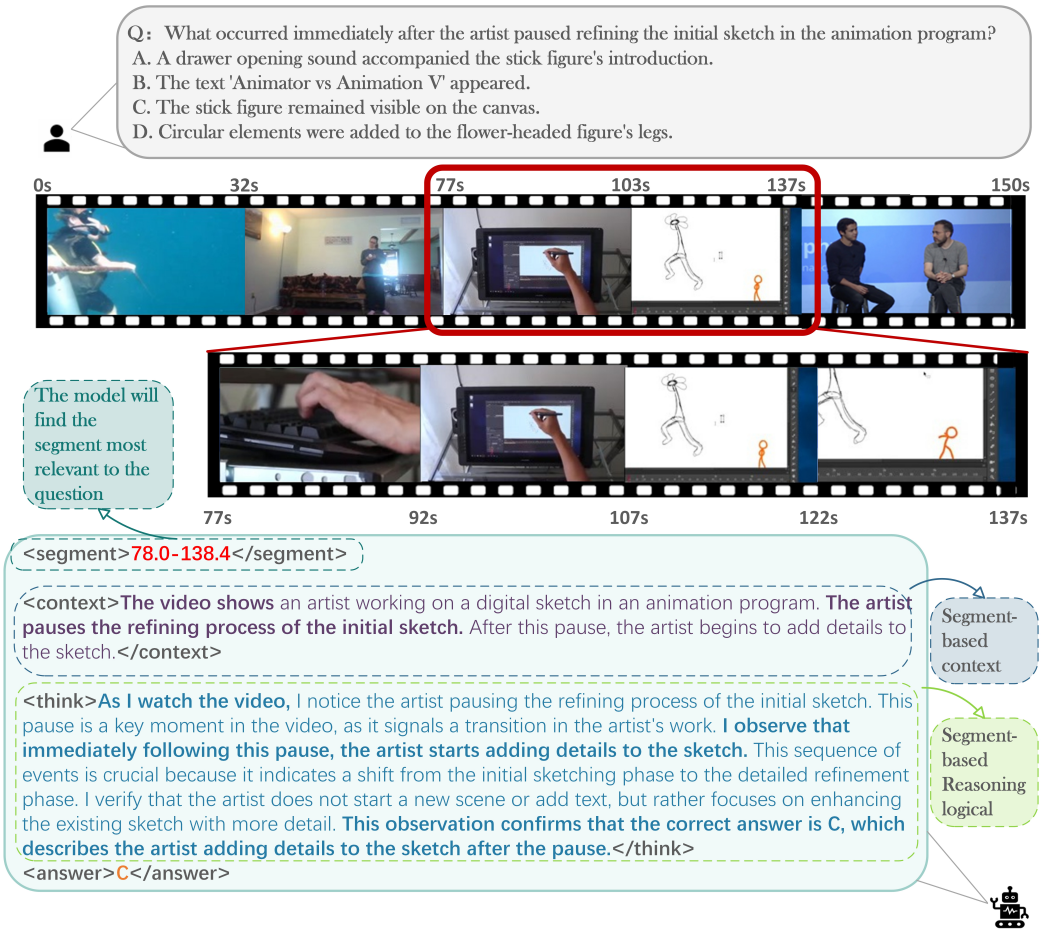

Q: What occurred immediately after the artist paused refining the initial sketch in the animation program?
A. A drawer opening sound accompanied the stick figure's introduction.
B. The text 'Animator vs Animation V' appeared.
C. The stick figure remained visible on the canvas.
D. Circular elements were added to the flower-headed figure's legs.
0s
32s
77s
103s
137s
150s
The model will find the segment most relevant to the question
77s
92s
107s
122s
137s
78.0-138.4
<context>The video shows an artist working on a digital sketch in an animation program. The artist pauses the refining process of the initial sketch. After this pause, the artist begins to add details to the sketch.</context>
Segment-based context
<think>As I watch the video, I notice the artist pausing the refining process of the initial sketch. This pause is a key moment in the video, as it signals a transition in the artist's work. I observe that immediately following this pause, the artist starts adding details to the sketch. This sequence of events is crucial because it indicates a shift from the initial sketching phase to the detailed refinement phase. I verify that the artist does not start a new scene or add text, but rather focuses on enhancing the existing sketch with more detail. This observation confirms that the correct answer is C, which describes the artist adding details to the sketch after the pause.</think>
Segment-based Reasoning logical
<answer>C</answer>

Figure 6: Reasoning path produced by our model on an example from ConcatBench. The model first identifies the part most relevant to the question and establishes the context of the video clip, then uses multimodal clues to answer the question accurately.

scenarios. We preserve all original QA annotations and automatically derive precise temporal segment labels during concatenation.

Filler Pool. To create temporal distractors, we build a filler pool from two sources. The first consists of external videos from additional datasets such as EMER (Lian et al., 2023) and Social-IQ (Zadeh et al., 2019), specified through YAML configuration. The second consists of mutual interference videos sampled from the available benchmark collections themselves.

Splicing Protocol. For each asked clip, we randomly sample $n_f \in [1, 4]$ filler clips from the pool and concatenate them into a single MP4 sequence using FFmpeg. The position of the asked clip is uniformly randomized within the sequence. The total duration of the concatenated video is strictly capped at 150 seconds. If the duration budget is exceeded, fewer filler clips are used. Each group contains 2 to 5 clips in total. This protocol is implemented by our automated pipeline.

Temporal Annotation. During concatenation, we record the exact start and end timestamps of every constituent clip via FFmpeg probe. The ground-truth segment interval of the asked clip is therefore derived automatically for each QA pair, providing temporal supervision without additional human annotation. The annotated interval is the boundary of the constituent clip that carries the question, not a minimal evidence interval within it. The metadata also stores per-clip durations, source provenance, and the complete segment interval sequence.

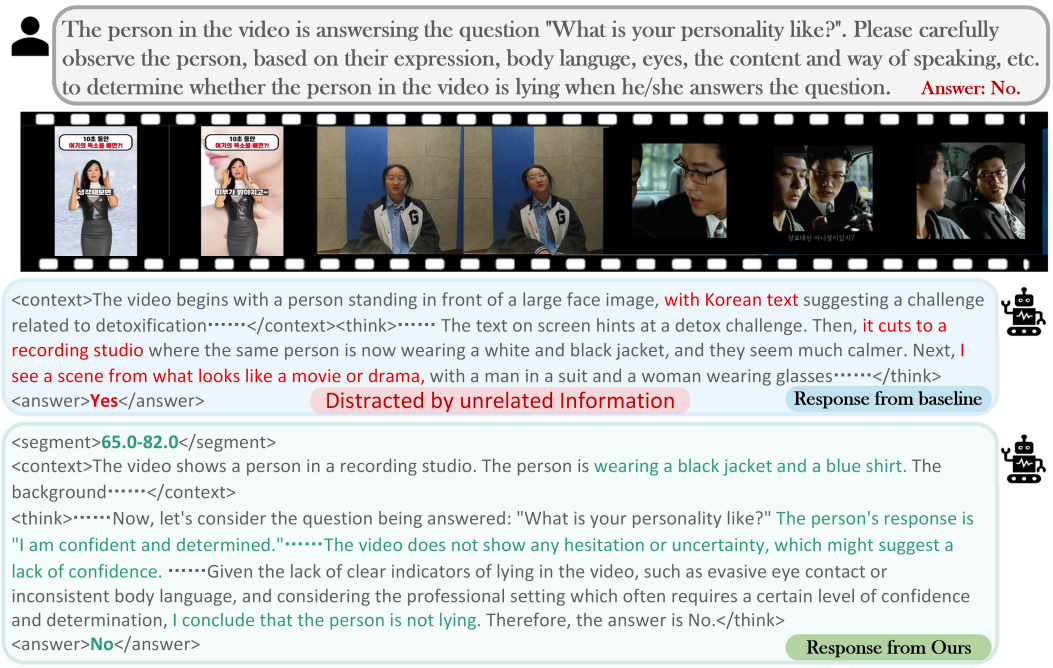

The person in the video is answersing the question "What is your personality like?". Please carefully observe the person, based on their expression, body languge, eyes, the content and way of speaking, etc. to determine whether the person in the video is lying when he/she answers the question. Answer: No.
<context>The video begins with a person standing in front of a large face image, with Korean text suggesting a challenge related to detoxification·····</context><think>····· The text on screen hints at a detox challenge. Then, it cuts to a recording studio where the same person is now wearing a white and black jacket, and they seem much calmer. Next, I see a scene from what looks like a movie or drama, with a man in a suit and a woman wearing glasses·····</think>
<answer>Yes</answer>
Distracted by unrelated Information
Response from baseline
65.0-82.0
<context>The video shows a person in a recording studio. The person is wearing a black jacket and a blue shirt. The background·····</context>
<think>·····Now, let's consider the question being answered: "What is your personality like?" The person's response is "I am confident and determined."·····The video does not show any hesitation or uncertainty, which might suggest a lack of confidence. ·····Given the lack of clear indicators of lying in the video, such as evasive eye contact or inconsistent body language, and considering the professional setting which often requires a certain level of confidence and determination, I conclude that the person is not lying. Therefore, the answer is No.</think>
<answer>No</answer>
Response from Ours

Figure 7: Qualitative comparison under temporal distraction. Both the baseline and FocusOmni receive the same question and concatenated video. The baseline attends to irrelevant video content and predicts "Yes," although the ground-truth answer is "No." In contrast, FocusOmni first grounds the relevant segment $[65.0, 82.0]$ s, focuses its visual and auditory reasoning on the localized evidence, and produces the correct answer "No." This example shows how explicit temporal grounding reduces interference from unrelated video segments.

B.2 Dataset Statistics and Properties

ConcatBench comprises 1,000 concatenated videos and 2,946 QA pairs spanning 9 problem categories. Multiple choice has 1,774 samples at 60.2%, event sequencing has 233 at 7.9%, audio-visual event alignment has 185 at 6.3%, context understanding has 154 at 5.2%, judge has 145 at 4.9%, reasoning has 126 at 4.3%, inference has 119 at 4.0%, emotion recognition has 105 at 3.6%, and comparative reasoning has 105 at 3.6%. Figure 2(a) shows the composition of asked clips by source dataset and problem type.

Comparison With Original Benchmarks. The original IntentBench, WorldSense, and Daily-Omni datasets evaluate holistic video understanding within a single isolated clip. In contrast, ConcatBench requires the model to first identify which temporal segment contains the relevant evidence among multiple concatenated clips and then answer the question based on that localized segment. This property makes ConcatBench suitable for benchmarking segment-aware models such as FocusOmni.

Data Format. Each sample is distributed with two JSON files. qa.json stores the question, answer, options, and a pointer to the concatenated video. meta.json stores the segment intervals of all clips, the index of the asked clip, source dataset provenance, and per-clip durations. This structured design facilitates joint evaluation of answer accuracy and temporal localization IoU.

B.3 Train/Validation/Test Splits

All experiments in this paper use a fixed partition of ConcatBench created before any fine-tuning and held constant across every model and ablation. Figure 2(b) shows the resulting counts.

The partition is designed to prevent leakage between splits. QA pairs are grouped by the connected components induced by shared source clips, and each complete group is assigned to a single split.

Thus, no source clip asked about in the test split appears in the training or validation splits. Within this constraint, groups are packed greedily by QA count to approximate a 70/10/20 partition at the QA level while balancing the nine problem categories across splits, yielding the realized proportions in Figure 2(b). The split is fixed before SFT, test videos are held out from training, training uses only the training split, checkpoint selection uses the validation split, and every result reported in Section 5.2 is computed on the held-out test split.

C Extended Experimental Details and Source-Dataset Results

C.1 Experimental Setup

Base Model. We use HumanOmniV2 (Yang et al., 2025) as the base model, built on the Qwen2.5-Omni-7B architecture (Xu et al., 2025). The model is fine-tuned with LoRA SFT on the training split of ConcatBench to produce the four-tag output format.

Datasets. We conduct the main experiments on ConcatBench as described in Sec. 4, with 1,000 concatenated videos and 2,946 QA pairs, split according to the train/validation/test partition in Appendix B.3. All ConcatBench numbers reported in this paper are computed on the held-out test split after selecting the checkpoint on validation, except for the controlled localization intervention in Section 5.3 and the localization quality analysis in Appendix D, which both run over the full benchmark pool. Additional evaluations are performed on three public omni-modal benchmarks in their original, non-concatenated form, which are IntentBench with 633 videos and 2,689 QA pairs, WorldSense with 1,662 videos and 3,172 QA pairs, and Daily-Omni with 684 videos and 1,197 QA pairs.

Baselines. We compare FocusOmni against three baselines. HumanOmniV2 (Yang et al., 2025) uses uniform 8-frame sampling as the direct baseline. Qwen2.5-VL-7B (Qwen Team et al., 2025) and Qwen2-VL-2B (Wang et al., 2024) also use uniform 8-frame sampling. All baselines use fixed fps with $f = 1.0$ and video-only input. None is run in its published default configuration. In particular, HumanOmniV2 samples 32 frames by default, while the 8-frame setting used here ensures that every baseline receives the same visual budget. GLM-4.6V in Table 1 is a member of the GLM-V series, which we cite through the series technical report (GLM-V Team et al., 2025).

FocusOmni Configuration. At inference time, FocusOmni uses a 3×3 grid for Stage 1 and $N_f = 8$ adaptively sampled frames for Stage 2, placed according to Eq. 6 over the predicted window, with video resolution 896×896 and full omni-modal input consisting of video and audio. The Daily-Omni transfer experiment in Table 3 uses $N_f = 16$ dense frames, while every other result uses $N_f = 8$. The SFT checkpoint is loaded for all FocusOmni experiments.

Metrics. On ConcatBench, we report answer accuracy by exact match, mean segment IoU, and IoU at thresholds 0.3, 0.5, and 0.7. The source datasets lack temporal annotations, so we report answer accuracy only.

Training Details. LoRA rank $r = 16$, $\alpha = 32$, target modules $\{W_q, W_k, W_v, W_o\}$. Learning rate 1×10^{-5} with cosine schedule and 5% warmup. Per-device batch size 1, 8-step gradient accumulation. 2,000 steps.

C.2 Results on Source Datasets

Table 3 and Table 4 report transfer results on the original, non-concatenated source datasets. Because these benchmarks do not contain temporal distractors, they serve as transfer diagnostics. The controlled ConcatBench comparison remains the main evidence for the benefit of temporal grounding under temporal distraction.

Official leaderboard context as of 2026-08-26. The Daily-Omni and WorldSense project leaderboards report newer systems beyond the reproduced reference rows above. For example, Daily-Omni lists Qwen2.5-Omni-7B at 62.07% and Qwen3-Omni-30B-A3B-Instruct at 71.85%, while WorldSense lists Qwen2.5-Omni at 45.4%, video-SALMONN 2+ 7B at 50.9%, and Qwen3-Omni at 54.0%. These official entries are retained only as temporal context. Their public records do not provide a uniform frame or token budget or fully matched inference configuration, so they are not used as direct controlled comparisons with FocusOmni.

Table 3: Transfer results on Daily-Omni (Zhou et al., 2025) under the stated reproduction protocol. FocusOmni uses one 3×3 grid overview plus 16 adaptively sampled dense frames.

Methods	LLM Size	AV Event Alignment	Comparative	Context Understanding	Event Sequence	Inference	Reasoning	Avg
Proprietary MLLMs								
Gemini 2.0 Flash (Comanici et al., 2025)	-	62.18	73.28	63.73	63.72	76.62	75.43	67.84
Gemini 2.0 Flash Lite (Comanici et al., 2025)	-	55.04	64.89	58.03	54.25	74.03	72.00	61.32
Open-Source Video-Audio MLLMs								
Unified-IO-2 (Lu et al., 2024)	7B	25.63	31.30	26.42	25.82	35.06	29.71	28.24
VideoLLaMA2 (Cheng et al., 2024)	7B	35.71	35.88	35.75	31.70	40.91	34.29	35.17
Qwen2.5-Omni (Xu et al., 2025)	3B	38.66	48.09	33.68	33.99	54.55	44.00	40.52
Qwen2.5-Omni (Xu et al., 2025)	7B	44.12	51.15	38.86	40.52	57.79	61.71	47.45
Ola (Liu et al., 2025)	7B	40.33	60.30	39.89	44.11	61.03	66.28	49.87
MiniCPM-o (OpenBMB Team, 2025)	7B	40.33	61.06	49.22	48.36	68.83	61.14	53.13
FocusOmni (16 frames)	7B	**43.70**	**67.94**	**49.74**	**44.77**	**69.93**	**68.00**	**53.88**

Table 4: Transfer results on WorldSense (Hong et al., 2025) under the stated reproduction protocol. FocusOmni uses one 3×3 grid overview with 9 labeled thumbnails plus 8 adaptively sampled dense frames.

Methods	LLM Size	Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	Avg
Proprietary MLLMs										
Claude 3.5 Sonnet (Anthropic, 2024)	-	43.7	31.7	30.6	36.5	30.7	31.9	36.6	33.9	34.8
GPT-4o (Hurst et al., 2024)	-	48.0	44.0	38.3	43.5	41.9	41.2	42.6	42.7	42.6
Gemini 1.5 Pro (Gemini Team et al., 2024)	-	53.7	47.2	50.3	50.4	52.4	46.8	40.2	42.0	48.0
Open-Source Video-Audio MLLMs										
Unified-IO-2 (Lu et al., 2024)	7B	27.1	31.7	23.9	23.7	25.5	23.7	25.7	27.3	25.9
VideoLLaMA2 (Cheng et al., 2024)	7B	29.4	25.4	21.8	24.5	26.2	24.6	25.5	27.1	25.4
VITA-1.5 (Fu et al., 2025)	7B	38.2	35.9	34.3	**39.8**	41.2	32.6	34.7	39.9	36.9
FocusOmni (8 frames)	7B	**47.6**	**47.2**	**42.4**	39.6	**42.3**	**45.1**	**40.5**	**44.0**	**43.5**

D Additional Ablations and Analyses

D.1 Per-Problem-Type Analysis

Table 1 provides the complete per-category comparison. This appendix view is included for completeness. The main-text discussion in Sec. 5.2 already summarizes the principal pattern. The gains are most visible on categories that require temporal evidence integration.

D.2 Localization Quality Analysis

We further analyze the quality and impact of temporal localization. Figure 8 shows the distribution of per-sample segment IoU. Like the controlled localization experiment, this analysis runs over the full ConcatBench pool. It is not restricted to the held-out test split. Approximately 34% of samples have zero IoU. This population should not be equated with Stage 1 parsing failure. It includes samples in which the segment field was invalid or unparseable, causing the system to fall back to uniform sampling, and samples in which a well-formed segment was predicted but did not overlap the ground-truth constituent clip at all. Our current statistics do not separate these two cases, so we report the aggregate and treat the distinction as an open item. Among the successfully localized samples with IoU > 0, the distribution is right-skewed. 50.4% achieve IoU ≥ 0.3, 42.3% achieve IoU ≥ 0.5, and 31.1% achieve IoU ≥ 0.7, with a mean IoU of 0.582 within this subset. This indicates that when the grid produces a usable segment, localization tends to be reasonably precise.

D.3 Loss Masking Strategy

We conducted preliminary experiments with three loss masking strategies during SFT. The first applies loss only to segment tokens. The second uses structure routing, applying loss only to tag boundaries while masking content. The third applies loss to all tokens, and we adopt it as the default. The first two strategies both led to degenerate autoregressive generation, producing repetitive token sequences or tag fragments. Full-sequence training was the only strategy that reliably produced correct four-tag output.

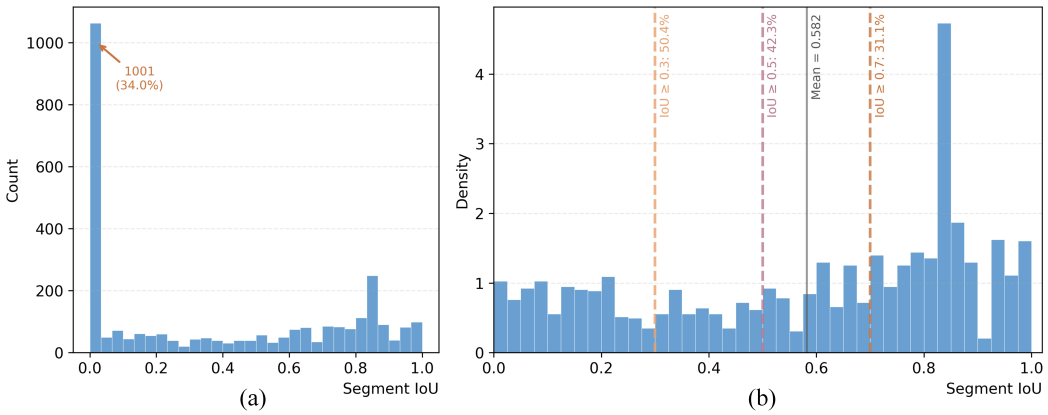


Figure 8: Distribution of segment IoU per sample. In both panels the horizontal axis is segment IoU. (a) The full distribution over all test samples, with the zero IoU count annotated. (b) The density over samples with IoU > 0, annotated with the fraction of samples above each threshold and the mean of this subset.

E LIMITATIONS

The controlled results establish the value of temporally relevant evidence within a fixed local-reasoning pipeline, but the two-stage design retains practical limitations. Stage 1 is vision-only, so events whose evidence is primarily auditory are localized from visual cues alone. Parsing failures trigger a fallback to uniform sampling, forfeiting the coarse-to-fine benefit after the extra Stage 1 pass has already been paid for, and our statistics do not separate such failures from ordinary localization error. The two-pass pipeline also adds latency. Our efficiency claim concerns visual input. Measured wall-clock per sample is ≈45% higher than single-pass 8-frame inference. Short events are difficult to localize, the grid resolution is fixed, and ConcatBench is synthesized by clip concatenation, which may introduce artifacts or discontinuities absent from natuFigure 4 focuses on the input-processingrally long videos. Distractors are not semantically filtered, so some may occasionally relate to the question, and the ground truth marks constituent-clip boundaries. It does not mark minimal evidence intervals. These limitations bound the scope within which our results should be interpreted.