Vidi2: Large Multimodal Models for Video Understanding and Creation

Intelligent Editing Team*, Intelligent Creation, ByteDance Inc.
San Jose/Seattle, US
https://bytedance.github.io/vidi-website/

Abstract

Video has emerged as the primary medium for communication and creativity on the Internet, driving strong demand for scalable, high-quality video production. Vidi models continue to evolve toward next-generation video creation and have achieved state-of-the-art performance in multimodal temporal retrieval (**TR**). In its second release, Vidi2 advances video understanding with fine-grained spatio-temporal grounding (**STG**) and extends its capability to video question answering (**Video QA**), enabling comprehensive multimodal reasoning. Given a text query, Vidi2 can identify not only the corresponding timestamps but also the bounding boxes of target objects within the output time ranges. This end-to-end spatio-temporal grounding capability enables potential applications in complex editing scenarios, such as plot or character understanding, automatic multi-view switching, and intelligent, composition-aware reframing and cropping. To enable comprehensive evaluation of STG in practical settings, we introduce a new benchmark, **VUE-STG**, which offers four key improvements over existing STG datasets: 1) **Video duration**: spans from roughly 10s to 30 mins, enabling long-context reasoning; 2) **Query format**: queries are mostly converted into noun phrases while preserving sentence-level expressiveness; 3) **Annotation quality**: all ground-truth time ranges and bounding boxes are manually annotated with high accuracy; 4) **Evaluation metric**: a refined vIoU/tIoU/vIoU-Intersection scheme for multi-segment spatio-temporal evaluation. In addition, we upgrade the previous VUE-TR benchmark to **VUE-TR-V2**, achieving a more balanced video-length distribution and more user-style queries. Remarkably, the Vidi2 model substantially outperforms leading proprietary systems, such as Gemini 3 Pro (Preview) and GPT-5, on both VUE-TR-V2 and VUE-STG, while achieving competitive results with popular open-source models with similar scale on video QA benchmarks.

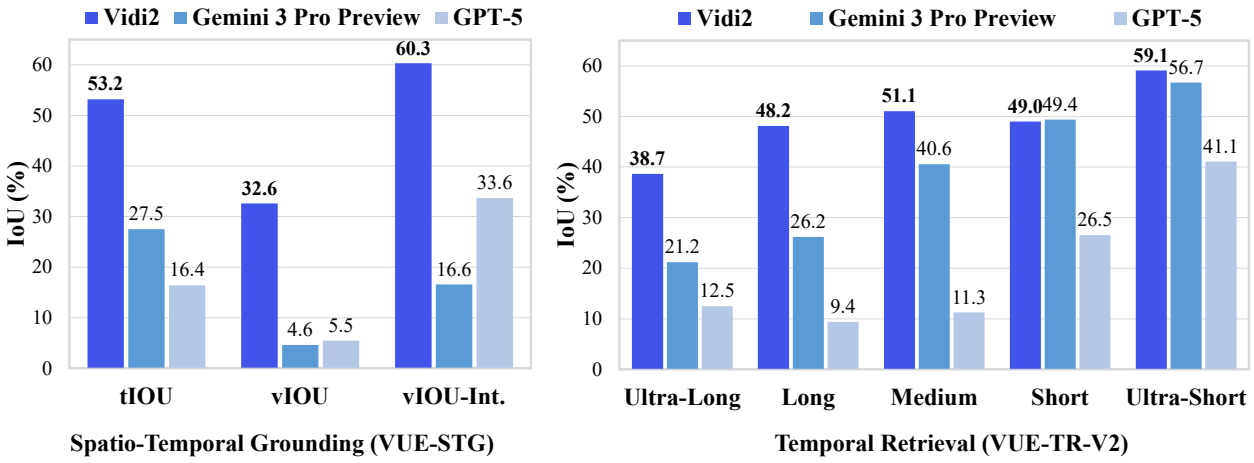


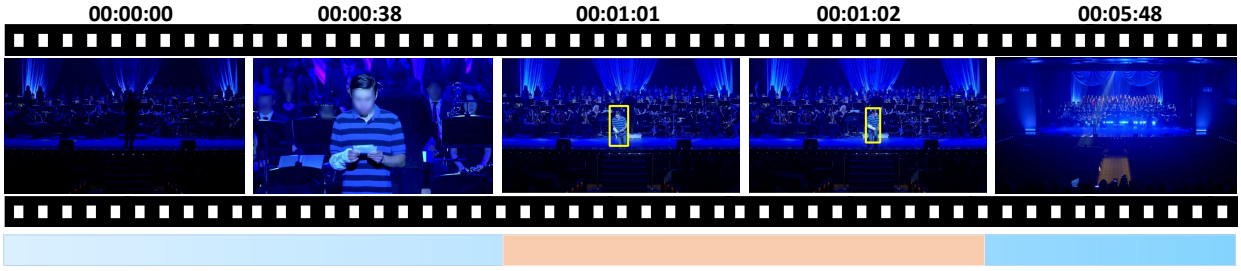
Figure 1: Spatio-temporal grounding and temporal retrieval results on the proposed benchmarks.

*A detailed contributor list can be found in Section 7.

1 Introduction

As an inherently multimodal medium, video has become a dominant channel for online communication and information sharing. However, producing high-quality videos remains challenging for most users, particularly when performing trimming or editing operations on mobile devices. The first release of **Vidi** [14] demonstrated strong temporal understanding across text, visual, and audio modalities, revealing its potential for automatic video trimming and understanding. To advance toward next-generation video creation systems capable of handling complex editing scenarios, Vidi models have been continually evolving to support more comprehensive multimodal perception and reasoning capabilities.

The second release, **Vidi2**, introduces for the first time an end-to-end spatio-temporal grounding (STG) capability, identifying not only the temporal segments corresponding to a text query but also the spatial locations of relevant objects within those frames. Notably, existing academic models [5, 6, 7, 8, 10, 16] and state-of-the-art industrial systems such as Gemini 3 Pro (Preview) [1, 3] and GPT-5 [9] are not yet capable of producing such fine-grained grounding results in a unified text output format (see Fig. 1). As illustrated in Fig. 2, the query "a man standing up from a kneeling position" represents a particularly difficult case involving multiple people in a dark scene. Vidi2 accurately localizes the corresponding time ranges and distinguishes the target person from others through precise bounding boxes in an end-to-end manner. It requires a comprehensive understanding of the visual and temporal dynamics of the described action. This fine-grained spatio-temporal perception highlights Vidi2's potential to enable advanced editing workflows, such as plot-level understanding, automatic multi-view switching, and composition-aware reframing for professional video creation.



Query (Vision, sentence): The man who is standing up from a kneeling position.

Predicted Tubes:
00:01:01: (0.452, 0.355, 0.530, 0.652)
00:01:02: (0.491, 0.371, 0.551, 0.627)
00:01:03: (0.491, 0.285, 0.562, 0.633)

Ground-truth Tubes:
00:01:01: (0.462, 0.391, 0.537, 0.657)
00:01:02: (0.494, 0.375, 0.557, 0.659)
00:01:03: (0.494, 0.281, 0.569, 0.660)

Figure 2: Examples of spatio-temporal grounding queries and their corresponding time ranges and object tubes (timestamps with bounding boxes, shown in yellow). Bounding boxes are expressed in percentage coordinates. The example video has a total duration of 387 seconds (*i.e.*, 06:27), and the query has been converted into a noun-style format. Facial regions are blurred to protect privacy.

To enable comprehensive evaluation of spatio-temporal grounding (STG) in realistic scenarios, we introduce a new benchmark, VUE-STG, featuring videos that range from 10 seconds to 30 minutes, substantially longer than existing STG datasets [2, 11, 18, 19]. Compared to prior work, VUE-STG includes videos and annotated tubes of much greater temporal extent, along with more complex, multimodal queries that often incorporate audio cues to assist temporal localization. All timestamps and bounding boxes are manually annotated with high precision, ensuring reliable spatio-temporal alignment. In addition, we upgrade the previous VUE-TR benchmark to VUE-TR-V2, featuring more human-authored, free-form queries and a more balanced video-length distribution with an increased proportion of long and ultra-long videos, making the task more realistic and challenging.

On the VUE-STG benchmark, Vidi2 surpasses leading proprietary models (Gemini 3 Pro (Preview) and GPT-5) by a substantial margin, demonstrating state-of-the-art spatio-temporal grounding capability among large multimodal models. Compared with the previous version, Vidi2 also shows significant improvements

in temporal retrieval, particularly for long videos. On the updated VUE-TR-V2 benchmark, it outperforms Gemini 3 Pro (Preview)[1] and GPT-5 by a wide margin across "Medium" to "Ultra-Long" video categories, confirming its leading performance in temporal reasoning. Furthermore, Vidi2 extends its capability to generic video question answering, marking a step toward a foundation model for comprehensive video understanding with strong spatio-temporal perception.

2 Model Overview

Vidi2 builds upon the foundation of Vidi [14] with substantial upgrades in architecture, training data, and task capabilities, most notably the support for spatio-temporal grounding (STG) and video question answering (Video QA). These improvements collectively enable fine-grained multimodal understanding across text, visual, and audio modalities.

Architecture: Vidi2 retains the multimodal architecture of Vidi [14], designed to jointly process text, visual, and audio inputs, while introducing key enhancements in both the encoder and LLM backbone (*e.g.*, Gemma-3 [12]). All results presented in this report are based on the 12B-parameter configuration. To achieve robust performance across videos of different lengths, we re-design the adaptive token compression strategy, improving the balance between short and long video representation efficiency. Additionally, a single image is treated as a 1-second silent video, ensuring a unified encoding interface for both image and video inputs.

Training Data: Vidi2 follows a training pipeline similar to the previous Vidi version but scales up in both data volume and data diversity. While synthetic data remains essential for coverage and stability, increasing the proportion of real video data significantly enhances performance across all video-related tasks. The supervised fine-tuning (SFT) dataset for temporal retrieval has been expanded and refined in both quality and quantity. Furthermore, we incorporate additional generic QA data during SFT to strengthen open-ended reasoning on images and videos, improving the model's versatility on downstream applications such as highlight detection, dense captioning, and video summarization.

Spatio-Temporal Grounding: To support the newly introduced spatio-temporal grounding (STG) capability, we introduce task-specific data across all training stages. Specifically, we leverage existing image-level spatial grounding datasets to synthesize large-scale spatio-temporal video grounding pairs, effectively bridging spatial and temporal alignment. In addition, we curate a substantial collection of real-world video STG annotations, which contribute significantly to the model's ability to localize both time ranges and bounding boxes with high precision.

Video Question Answering: The previous Vidi [14] model demonstrated the effectiveness of the temporal-aware multimodal alignment training stage (see Sec. 4 in [14]) for temporal retrieval. Vidi2 further validates the generalization capability of this multimodal alignment strategy to generic video QA. By incorporating video QA data during the post-training phase, Vidi2 is able to answer both visual and auditory questions with high accuracy. We also observe that scaling up multimodal alignment data consistently improves overall video QA performance.

3 Evaluation Benchmark

To comprehensively evaluate the performance and generalization of Vidi2, we consider three representative categories of video understanding tasks: Spatio-Temporal Grounding (STG), Temporal Retrieval (TR), and Video Question Answering (Video QA). The first two are evaluated on our newly proposed benchmarks (*i.e.*, VUE-STG and VUE-TR-V2), which are designed to assess fine-grained temporal and spatio-temporal reasoning under realistic editing and understanding scenarios. For video QA, we adopt widely used public benchmarks, including LVBench [15], LongVideoBench [17], and VideoMME [4]. All of these video QA datasets employ a multiple-choice question format, which enables standardized, objective, and reproducible

[1]Note that Gemini 3 Pro shows a notable improvement over Gemini-2.5-Pro-0325 on temporal retrieval.

evaluation of multimodal reasoning performance. Together, these datasets provide a unified and comprehensive evaluation protocol that spans retrieval, grounding, and video reasoning.

3.1 VUE-STG

We introduce **VUE-STG**, a new **V**ideo **U**nderstanding **E**valuation benchmark specifically designed to advance **S**patio-**T**emporal **G**rounding in real-world scenarios. VUE-STG addresses four critical aspects often overlooked in prior academic benchmarks [2, 11, 18, 19]: video duration, query format, annotation quality, and metric design. To ensure the accurate performance signal, all ground-truth labels are manually curated, which yields significantly more accurate and consistent labels than existing benchmarks.

Duration Category	Ultra-short < 1 min	Short $1 - 10$ mins	Medium $10 - 30$ mins	Total
# Videos	126	294	562	982
# Queries (Tubes)	206	461	933	1,600
# Boxes	1,063	3,741	7,343	12,147
Video Hours	0.82	26.28	177.69	204.79

Table 1: Video duration distribution of the proposed VUE-STG benchmark. The benchmark covers a wide range of video lengths, from ultra-short (< 1 min) clips to medium ($10 - 30$ mins) videos, enabling comprehensive evaluation of models across diverse real-world scenarios.

3.1.1 Data Distribution

Following VUE-TR [14], VUE-STG is built using publicly available videos, and the annotations are made accessible for open research. As presented in Table 1, the benchmark consists of $1,600$ queries across 982 videos, spanning over 204.79 hours. It supports attribute-based slicing for fine-grained performance analysis as follows.

- **Video Duration.** Unlike existing datasets limited to short videos, the video length varies from 10 seconds to 30 minutes, covering most of the duration ranges encountered in real-world scenarios. To enable duration-wise evaluation, we categorize videos into three balanced buckets following the definition of VUE-TR [14]: ultra-short (< 1 min), short ($1 - 10$ mins), and medium ($10 - 30$ mins).
- **Query.** Each query in VUE-STG is carefully verified by humans, which clearly indicates the target object described in the video, aiming to minimize potential ambiguity in grounding. When a descriptive sentence naturally involves multiple objects, we adjust the phrasing to better highlight the intended target. For example, instead of using an ambiguous expression such as "a player is loaded into an ambulance," we adopt more explicit formulations such as "the ambulance which the player is being loaded into" or "the player who is loaded into the ambulance," depending on which object is referenced. This annotation strategy substantially reduces semantic ambiguity and improves the reliability of query-object associations.
- **Tube.** Each target object is annotated as a spatio-temporal tube at 1 frame per second, ensuring consistent localization quality across long-form videos. Object-size statistics are computed based on the average bounding box area of each tube, resulting in a balanced distribution, as shown in Table 2. Furthermore, Table 3 summarizes the distribution of the temporal duration of the tubes. While most tubes span 3 to 10 seconds, both short transient events and long-duration activities are included to support a comprehensive evaluation. We note that not all tubes in VUE-STG are strictly temporally contiguous. In real-world scenarios, target objects may disappear due to camera cuts, occlusions, or transitions between scenes. To better reflect these conditions, we preserve such temporally fragmented tubes rather than enforcing artificial continuity. These fragmented cases allow models to be evaluated under more realistic long-form video dynamics, where objects may intermittently leave and re-enter the visible field.

Object Size	Small $< 10\%$	Medium $10\% - 30\%$	Large $> 30\%$
# Queries (Tubes)	746	511	343

Table 2: Object-size distribution of annotated tubes in the VUE-STG benchmark. Object size is defined as the average bounding-box area within each tube, normalized by the video frame area.

Tube Duration	Micro-Short $< 3s$	Ultra-Short $3s-10s$	Short $10s-60s$
# Queries (Tubes)	179	1013	408

Table 3: Temporal duration distribution of annotated tubes in the VUE-STG benchmark. Tubes cover short to long temporal extents, from brief moments (< 3s) to longer activities (10–60s), enabling fine-grained analysis of temporal grounding performance.

3.1.2 Evaluation Metrics

We evaluate both *temporal grounding accuracy* and *spatio-temporal grounding accuracy* to assess model performance. The temporal metrics measure how well the predicted time spans align with the ground-truth temporal annotations, whereas the spatio-temporal metrics evaluate the localization quality of predicted bounding boxes across time. All results are reported as sample-wise averages on the VUE-STG test set. Among these, vIoU serves as the primary ranking metric, jointly capturing both temporal and spatial alignment accuracy.

Tube representation. A spatio-temporal tube is modeled as a time-dependent bounding box, *i.e.*,

$$B(t) = \Big(x_0(t),\, y_0(t),\, x_1(t),\, y_1(t)\Big), \tag{1}$$

where $(x_0(t), y_0(t))$ and $(x_1(t), y_1(t))$ denote the top-left and bottom-right coordinates. Let $B^{\text{pred}}(t)$ and $B^{\text{gt}}(t)$ denote the predicted and ground-truth tubes, defined on temporal supports $T_{\text{pred}}, T_{\text{gt}} \subset \mathbb{R}$. In practice, the temporal axis is discretized by uniform sampling, *i.e.*, we use the sampling rate 1 frame per second throughout this work.

Bounding-box IoU (bIoU**).** We define

$$\text{bIoU}(B_1, B_2) = \frac{\text{Area}(B_1 \cap B_2)}{\text{Area}(B_1 \cup B_2)} \tag{2}$$

as the spatial overlap metric between two bounding boxes B_1 and B_2.

Temporal metrics: Temporal grounding metrics evaluate how well the predicted time interval aligns with the ground-truth annotated temporal extent, including tIoU, tP and tR. Let $T_\cap = T_{\text{pred}} \cap T_{\text{gt}}$ and $T_\cup = T_{\text{pred}} \cup T_{\text{gt}}$. Temporal precision and recall (tP and tR) measures how much of the predicted time span overlaps with the ground truth, while temporal recall measures how much of the ground truth duration is covered by the prediction:

$$\text{tP} = \frac{|T_\cap|}{|T_{\text{pred}}|}, \qquad \text{tR} = \frac{|T_\cap|}{|T_{\text{gt}}|}.$$

Meanwhile, temporal IoU (tIoU) measures the alignment between the predicted and ground-truth time intervals using an IoU-style ratio:

$$\text{tIoU} = \frac{|T_\cap|}{|T_\cup|}.$$

Spatio-temporal metrics: Spatio-temporal metrics evaluate how well the model localizes the target object throughout the video, including vIoU, vIoU-Int., vP, and vR.

Frame-level IoU. For any $t \in T_{\text{pred}} \cup T_{\text{gt}}$, we define the frame-level IoU as

$$\text{IoU}_t = \begin{cases} \text{bIoU}\big(B^{\text{pred}}(t),\, B^{\text{gt}}(t)\big), & t \in T_{\cap}, \\ 0, & \text{otherwise}, \end{cases}$$

where $T_{\cap} = T_{\text{pred}} \cap T_{\text{gt}}$ is the temporal intersection.

Accumulated IoU over the temporal intersection. Since $\text{IoU}_t = 0$ outside $T_{\cap}$, the accumulated IoU over any time domain reduces to the sum over $T_{\cap}$. We define

$$S = \sum_{t \in T_{\cap}} \text{IoU}_t.$$

Spatio-temporal precision and recall (vP and vR). We measure the average frame-level IoU over the predicted and ground-truth temporal spans:

$$\text{vP} = \frac{S}{|T_{\text{pred}}|}, \qquad \text{vR} = \frac{S}{|T_{\text{gt}}|}.$$

Spatial IoU over the temporal union (vIoU). To evaluate spatio-temporal accuracy over the entire time span where either tube exists, we average the frame-level IoU over the temporal union $T_{\cup} = T_{\text{pred}} \cup T_{\text{gt}}$:

$$\text{vIoU} = \frac{S}{|T_{\cup}|}.$$

Spatial IoU over the temporal intersection (vIoU-Int.). For completeness, we also compute the average IoU over the time interval where both tubes are defined as

$$\text{vIoU-Int.} = \frac{S}{|T_{\cap}|}.$$

3.2 VUE-TR-V2

Following the settings of VUE-TR [14], we introduce an updated version, VUE-TR-V2, to provide a more balanced video length distribution and better query quality of the evaluation dataset.

Data Distribution: As shown in Table 4, VUE-TR-V2 contains a comparable number of videos to VUE-TR, but the total duration increases substantially from 107.87 hours to 311.11 hours. The modality and format distributions (Figure 4) remain consistent with the original benchmark. To better reflect real-world conditions, VUE-TR-V2 introduces a greater proportion of long and ultra-long videos, resulting in a more balanced length distribution where short and medium clips no longer dominate the dataset. It covers diverse practical scenarios such as movie mashups and long-form content.

Evaluation Metric: We follow the previous VUE-TR [14] benchmark that uses the AUC (Area Under Curve) score of temporal-dimension precision, recall, and IoU (Intersection over Union) in the evaluation, denoted as $\bar{P}$, $\bar{R}$, $\bar{\text{IoU}}$ in Sec. 4. In particular, $\bar{\text{IoU}}$ is used as the primary metric.

4 Experiment

To validate the effectiveness of Vidi2, we evaluate it on three representative tasks: Temporal Retrieval, Spatio-Temporal Grounding, and Video Question Answering. While the first two are tested on our proposed benchmarks (VUE-TR-V2 and VUE-STG), the latter leverages public video QA datasets [15, 17, 4] to assess general multimodal understanding.

Duration Category	Ultra-short < 1 min	Short $1 - 10$ mins	Medium $10 - 30$ mins	Long $30 - 60$ mins	Ultra-long > 60 mins	Total
# Videos	66	230	315	197	39	847
# Queries	134	390	430	426	220	1,600
Video Hours	0.74	19.85	97.08	145.95	47.50	310.72

Table 4: Duration distribution of videos in the proposed VUE-TR-V2 evaluation benchmark. The dataset covers a wide range of video lengths, from ultra-short clips ($<$ 1 minute) to ultra-long videos ($>$ 1 hour), enabling comprehensive evaluation of temporal retrieval models across diverse real-world scenarios. The total video length (310.72 hours) of VUE-TR-V2 is increased to over 2.8 times of the video length (107.87 hours) in the previous VUE-TR benchmark.

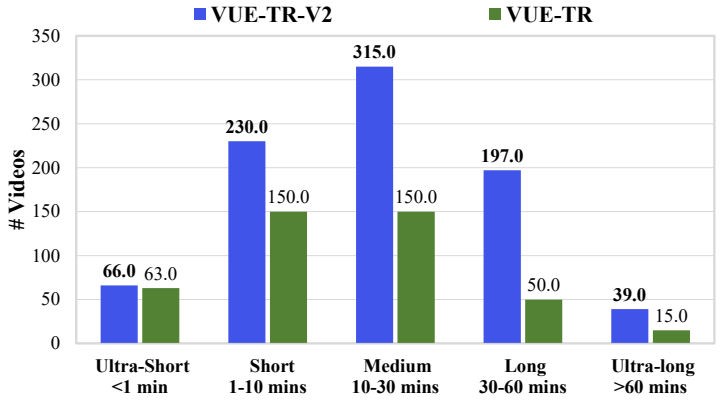


Figure 3: The video distribution comparison between the VUE-TR-V2 and VUE-TR benchmarks.

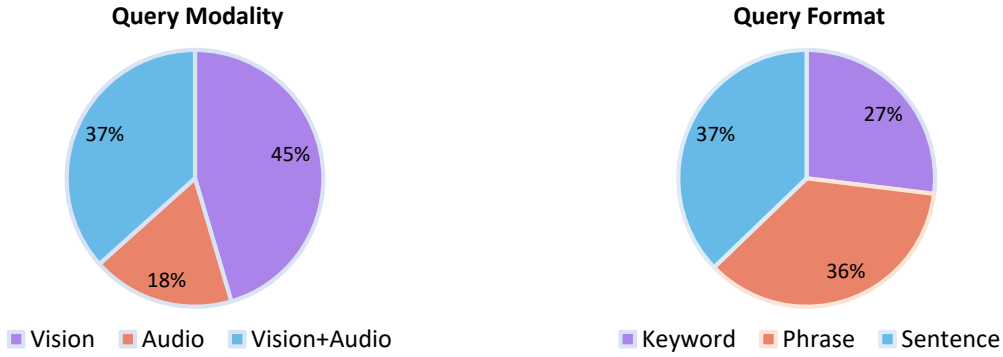


Figure 4: The distribution of query modality and format in the VUE-TR-V2 benchmark.

4.1 Spatio-Temporal Grounding

We compare our Vidi2 model against leading large multimodal models (LMMs), including Gemini 3, GPT-5 and Qwen3-VL, on the VUE-STG benchmark. Since our system predicts full spatio-temporal tubes, we conduct a comprehensive analysis across both temporal and spatial dimensions. The metrics follow the definitions in Sec. 3.1.2.

Overall, Vidi2 achieves the strongest performance across all temporal and spatio-temporal metrics, consistently outperforming Gemini 3, GPT-5 and Qwen3-VL. This demonstrates that Vidi2 not only produces more accurate temporal grounding but also achieves substantially better spatial localization throughout the video sequence.

4.1.1 Inference Setup for Multimodal LLM Baselines

To ensure a fair comparison with other LMMs, we carefully design task-aligned prompts and standardized output formats so that all models can perform spatio-temporal grounding in a one-shot setting. Our Vidi2 model outputs spatio-temporal tubes as a sequence of discrete, normalized bounding boxes per timestamp, while other LMMs generate temporal predictions in varying forms according to their APIs. In the following, we detail the input, prompt, and output settings for each model, and describe how all predictions are converted into a unified temporal representation for consistent evaluation.

Gemini 3 Input/Output Format: We evaluate the Gemini 3 Pro (Preview) model, which accepts video input via URL. Specifically, we design a structured prompt that instructs the model to output a JSON array containing timestamps and bounding boxes, *i.e.*,

```
Answer with timestamps and bounding boxes and do not output explanation.
Output only a JSON array.
Timestamp format: MM:SS with zero-padding (00-59 for MM and SS).
Bounding box format: [x0, y0, x1, y1] with values in [0, 1000] normalized to the video frame.

Example:
[{"timestamp":"00:30", "box_2d":[100, 200, 300, 400]},
 {"timestamp":"05:00", "box_2d":[150, 250, 350, 450]}]

What are all the timestamps and positions corresponding to the text query:
"yellow school bus which parked on street near the tree with yellow leaves"?
```

In constructing our prompts, we follow the official Gemini API guidelines for temporal and spatial annotations: the `MM:SS` timestamp format is based on the Video Understanding specification [2], while the bounding box coordinate system in $[0, 1000]$ is defined in the Image Understanding documentation [3].

GPT-5 Input/Output Format: GPT-5 API does not accept video directly; instead, it supports a sequence of up to 120 image frames per request. We sample frames according to the following rule: (1) for videos shorter than two minutes, we extract frames at 1 FPS, matching Vidi2's input configuration; (2) for longer videos, we uniformly subsample frames to satisfy the 120-frame limit. The prompt for GPT-5 is

```
Given the frames of video, please find all the objects corresponding to the text query:
"yellow school bus which parked on street near the tree with yellow leaves".

Output only a JSON array.
Frame index format: 0-based integer.
Bounding box format: [x0, y0, x1, y1] with normalized value in 0.000 ~ 1.000

Example:
[{"frame": 3, "box": [0.051, 0.252, 0.323, 0.954]},
{"frame": 5, "box": [0.372, 0.353, 0.634, 0.955]}]
```

The model outputs bounding boxes defined on sampled frame indices.

Qwen3-VL Input/Output Format: For Qwen3-VL, we test the `Qwen3-VL-32B-Instruct` model following the spatio-temporal grounding examples provided in the Qwen3-VL blog[4] and the interleaved timestamp-image pattern demonstrated in the official cookbook[5].
Following the conventions illustrated by the cookbook, each video is converted into a sequence of sampled frames, and the prompt is formed by interleaving explicit timestamp indicators (*e.g.*, "`<0.0 seconds>`"),

[2] `https://ai.google.dev/gemini-api/docs/video-understanding`
[3] `https://ai.google.dev/gemini-api/docs/image-understanding`
[4] `https://qwen.ai/blog?id=qwen3-vl`
[5] `https://github.com/QwenLM/Qwen3-VL/blob/main/cookbooks/video_understanding.ipynb`

the corresponding frame image, the next timestamp, and so on. After all timestamp-image pairs are provided, the textual grounding query is appended, *e.g.*,

```
Given the query "yellow school bus which parked on street near the tree with
yellow leaves", for each frame, detect and localize the visual content described
by the given textual query in JSON format. If the visual content does not exist in
a frame, skip that frame.

Output Format: [{"time": 1.0, "bbox_2d": [x_min, y_min, x_max, y_max]},
{"time": 2.0, "bbox_2d": [x_min, y_min, x_max, y_max]}, ...].
```

Qwen3-VL outputs timestamps directly in seconds, and its bounding box coordinates are defined over an integer range of $[0, 1000]$, which coincides with the range used by the Gemini API.

Temporal and Spatial Normalization: The four models generate temporal grounding signals in mutually incompatible representations. To enable fair comparison, we convert all predictions into a unified timeline measured in seconds. Specifically, 1) Gemini 3's `MM:SS` timestamps are parsed and converted into seconds; 2) GPT-5's frame indices are mapped to seconds according to the sampling policy; 3) Qwen3-VL directly reports timestamps in seconds, requiring no additional temporal transformation. After conversion, all timestamps are discretized by rounding to the nearest integer second to obtain a consistent per-second temporal representation.
For spatial normalization, we map all bounding boxes to a shared coordinate system in $[0, 1]$. Gemini and Qwen3-VL use an integer-based coordinate range of $[0, 1000]$, which we normalize by dividing by 1000. GPT-5 already produces normalized coordinates. This yields a unified spatial representation from which all spatial and spatio-temporal metrics are computed.

4.1.2 Performance Evaluation

Temporal Performance: Table 5 summarizes the temporal grounding results across different video lengths, ground-truth temporal durations, and object sizes. Vidi2 establishes a clear advantage in all conditions, achieving the highest temporal precision (73.00), recall (59.80), and IoU (53.19). The gap is especially pronounced on medium-length videos, where Gemini 3 Pro (Preview), GPT-5 and Qwen3-VL-32B degrade significantly. These findings highlight that Vidi2 maintains robust temporal alignment even under long-duration or fine-grained temporal conditions.

Spatio-Temporal Performance: Table 6 reports the spatio-temporal grounding accuracy. Across all categories, Vidi2 again delivers the strongest results, with the highest vP (44.56), vR (36.32), and vIoU (32.57), as well as a large margin on the intersection-based metric vIoU-Int (60.30). These improvements indicate that Vidi2 not only identifies the correct temporal extent but also maintains coherent spatial localization throughout the entire tube. In contrast, Gemini 3 Pro (Preview), GPT-5 and Qwen3-VL-32B exhibit severe degradation on long videos or small-object scenarios, suggesting instability in spatial tracking when reasoning over extended temporal horizons.

4.2 Temporal Retrieval

We compare our model with three state-of-the-art proprietary models including Gemini 3 Pro (Preview) [3] and GPT-5 [9]. There models are chosen for their strong performance and broad modality support, making them competitive baselines for real-world temporal retrieval.

The latest Gemini 3 Pro (Preview) naturally supports '$*$.mp4' as video input with maximum length over 2 hours. Compared with previous 0325 version, the latest version can follow the temporal retrieval task very well and the result stably contains timestamps in `HH:MM:SS`. One successful example prompt is used as follows.

```
Answer with time ranges and do not output explanation.  What are all the time
ranges corresponding to the text query:  "QUERY"?.
```

Major Type	Category	Metric(%)	**Vidi2**	Gemini 3 Pro Prev.	GPT-5	Qwen3-VL-32B
Overall	Overall	tIoU	**53.19**	27.50	16.40	25.91
		tP	**73.00**	51.91	38.29	45.29
		tR	**59.80**	35.26	19.53	39.19
Video Length	Ultra-Short ($< 1m$)	tIoU	61.43	38.68	**64.66**	54.33
		tP	**85.04**	58.91	78.34	76.54
		tR	66.80	52.29	**75.80**	69.22
	Short ($1 - 10m$)	tIoU	**61.48**	35.39	19.73	38.97
		tP	**81.83**	66.49	43.86	58.66
		tR	**68.36**	43.59	24.27	57.27
	Medium ($10 - 30m$)	tIoU	**47.27**	21.13	4.10	13.19
		tP	**65.98**	43.15	23.86	30.45
		tR	**54.01**	27.37	4.77	23.64
Tube Duration	Micro-Short ($< 3s$)	tIoU	**47.24**	25.92	16.70	24.44
		tP	**56.49**	37.06	30.61	44.74
		tR	**59.78**	39.66	20.95	35.75
	Ultra-Short ($3s-10s$)	tIoU	**54.66**	29.51	17.63	25.80
		tP	**73.13**	51.77	38.17	46.23
		tR	**60.85**	38.03	21.22	40.21
	Short ($10s-60s$)	tIoU	**52.14**	23.20	13.23	26.85
		tP	**79.93**	58.71	41.25	43.23
		tR	**57.18**	26.43	14.72	38.18
Object Size	Small ($< 10\%$)	tIoU	**51.46**	28.35	17.57	26.01
		tP	**70.79**	52.00	40.94	46.46
		tR	**57.94**	37.16	20.46	38.99
	Medium ($10\% - 30\%$)	tIoU	**54.51**	27.23	16.66	26.85
		tP	**73.85**	51.58	36.91	45.04
		tR	**61.56**	34.41	20.04	40.76
	Large ($> 30\%$)	tIoU	**54.97**	26.06	13.49	24.32
		tP	**76.55**	52.23	34.77	43.18
		tR	**61.21**	32.38	16.76	37.29

Table 5: Temporal grounding results on the VUE-STG benchmark, reporting temporal precision (tP), recall (tR), and intersection-over-union (tIoU). tR and tIoU are averaged over ground-truth tubes, while tP is averaged over predicted tubes. Results are further grouped by video length, annotated tube duration, and object size.

Major Type	Category	Metric(%)	**Vidi2**	Gemini 3 Pro Prev.	GPT-5	Qwen3-VL-32B
Overall	Overall	vIoU	**32.57**	4.61	5.47	5.12
		vIoU-Int.	**60.30**	16.59	33.64	18.47
		vP	**44.56**	8.95	13.01	8.61
		vR	**36.32**	5.71	6.50	7.49
Video Length	Ultra-Short ($< 1m$)	vIoU	**36.70**	7.77	20.60	11.30
		vIoU-Int.	**57.86**	19.44	31.36	20.56
		vP	**50.06**	12.44	25.19	16.11
		vR	**39.46**	9.96	23.98	13.95
	Short ($1 - 10m$)	vIoU	**39.62**	6.36	6.86	8.63
		vIoU-Int.	**62.97**	18.45	34.45	21.01
		vP	**52.41**	12.36	14.97	12.37
		vR	**43.45**	7.61	8.43	12.56
	Medium ($10 - 30m$)	vIoU	**28.18**	3.05	1.44	2.02
		vIoU-Int.	**59.37**	14.33	34.52	15.08
		vP	**39.46**	6.49	8.47	4.75
		vR	**32.10**	3.84	1.68	3.55
Tube Duration	Micro-Short ($< 3s$)	vIoU	**27.96**	3.93	5.60	4.31
		vIoU-Int.	**57.28**	12.98	32.95	16.06
		vP	**33.25**	5.85	10.52	8.10
		vR	**34.74**	5.20	6.74	5.34
	Ultra-Short ($3s - 10s$)	vIoU	**34.01**	4.79	5.74	4.57
		vIoU-Int.	**60.98**	16.30	33.40	17.32
		vP	**45.17**	8.66	12.90	8.07
		vR	**37.52**	6.08	6.94	7.15
	Short ($10s - 60s$)	vIoU	**31.03**	4.48	4.76	6.85
		vIoU-Int.	**59.83**	18.43	34.27	22.45
		vP	**48.00**	11.03	14.14	10.16
		vR	**34.01**	5.01	5.29	9.27
Object Size	Small ($< 10\%$)	vIoU	**23.31**	2.33	3.66	1.49
		vIoU-Int.	**44.13**	7.90	21.38	5.88
		vP	**32.08**	4.36	9.06	2.73
		vR	**25.71**	2.85	4.18	2.17
	Medium ($10\% - 30\%$)	vIoU	**39.42**	5.83	7.34	7.08
		vIoU-Int.	**71.21**	20.41	42.19	23.65
		vP	**52.97**	10.92	15.93	11.29
		vR	**44.44**	7.16	8.69	10.20
	Large ($> 30\%$)	vIoU	**42.50**	7.77	6.61	10.09
		vIoU-Int.	**77.22**	30.22	48.08	38.02
		vP	**59.14**	15.86	16.96	17.08
		vR	**47.29**	9.78	8.27	15.00

Table 6: Spatio-temporal grounding results on the VUE-STG benchmark, reporting spatio-temporal precision (vP), recall (vR), and intersection-over-union (vIoU and vIoU-Int.). vR and vIoU are averaged over ground-truth tubes, while vP is averaged over predicted tubes. vIoU-Int. is computed only over the predicted tubes that overlap with a ground-truth tube. Results are grouped by video length, annotated tube duration, and object size.

Category	Metric(%)	**Vidi2**	Gemini 3 Pro Prev.	GPT-5
Overall	$\bar{\text{IoU}}$	**48.75**	37.58	17.15
	$\bar{\text{P}}$	**62.45**	48.61	29.64
	$\bar{\text{R}}$	**64.93**	56.30	26.63
Ultra-Short ($< 1m$)	$\bar{\text{IoU}}$	**59.09**	56.72	41.08
	$\bar{\text{P}}$	**69.76**	68.56	63.84
	$\bar{\text{R}}$	**82.48**	75.22	48.19
Short ($1 - 10m$)	$\bar{\text{IoU}}$	49.01	**49.40**	26.52
	$\bar{\text{P}}$	**65.38**	61.83	41.64
	$\bar{\text{R}}$	66.05	**74.93**	38.91
Medium ($10 - 30m$)	$\bar{\text{IoU}}$	**51.05**	40.56	11.28
	$\bar{\text{P}}$	**63.42**	50.48	20.51
	$\bar{\text{R}}$	**67.91**	63.69	21.50
Long ($30 - 60m$)	$\bar{\text{IoU}}$	**48.15**	26.20	9.39
	$\bar{\text{P}}$	**60.54**	36.73	20.21
	$\bar{\text{R}}$	**62.56**	39.36	16.05
Ultra-Long ($> 60m$)	$\bar{\text{IoU}}$	**38.65**	21.19	12.49
	$\bar{\text{P}}$	**54.55**	32.18	21.49
	$\bar{\text{R}}$	**50.98**	30.11	22.27
Keyword	$\bar{\text{IoU}}$	**47.73**	38.41	23.17
	$\bar{\text{P}}$	**63.21**	52.43	41.07
	$\bar{\text{R}}$	**63.72**	56.00	32.59
Phrase	$\bar{\text{IoU}}$	**50.11**	37.84	17.02
	$\bar{\text{P}}$	**63.80**	48.94	29.18
	$\bar{\text{R}}$	**67.07**	57.40	25.88
Sentence	$\bar{\text{IoU}}$	**48.17**	36.73	12.92
	$\bar{\text{P}}$	**60.59**	45.51	22.26
	$\bar{\text{R}}$	**63.75**	55.46	23.04
Audio	$\bar{\text{IoU}}$	**46.90**	33.27	6.93
	$\bar{\text{P}}$	**57.51**	39.42	12.16
	$\bar{\text{R}}$	**67.09**	58.26	23.21
Vision	$\bar{\text{IoU}}$	**51.04**	39.54	20.62
	$\bar{\text{P}}$	**63.12**	52.16	35.78
	$\bar{\text{R}}$	**66.94**	54.76	26.85
Vision+Audio	$\bar{\text{IoU}}$	**46.81**	37.26	17.85
	$\bar{\text{P}}$	**64.03**	48.66	30.17
	$\bar{\text{R}}$	**61.38**	57.26	28.04

Table 7: Performance of different models on the VUE-TR-V2 benchmark across various evaluation attributes. $\bar{P}$ and $\bar{R}$ denote the AUC (Area Under Curve) of precision and recall, respectively, while $\bar{\text{IoU}}$ represents the AUC of intersection-over-union between predicted and ground-truth timestamp ranges, following the definition in [14]. GPT models are evaluated via the Azure API, and Gemini models via the official Google API. The latest Gemini 3 Pro (Preview) model is tested by directly uploading full videos (no size restriction), whereas GPT-5 is limited by the Azure API's 120-frame cap; for videos exceeding 120 seconds, we uniformly sample 120 frames.

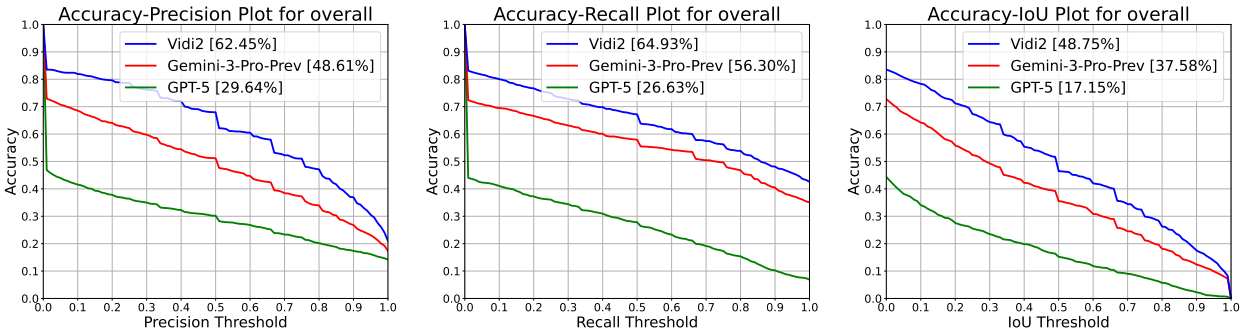


Figure 5: Overall performance curves for temporal retrieval on the VUE-TR-V2 benchmark. We report accuracy across varying thresholds for different models.

	LVBench [15]	LongVideoBench [17]	VideoMME [4]
Gemini-2.5-Pro [3]	78.7	84.3	-
Qwen2.5 VL-7B [13]	45.3	54.7	65.1
Vidi2	45.8	57.1	63.5

Table 8: Video QA performance of Vidi2 on popular benchmarks.

Since The lastest GPT-5 API does not support audio, we extract frames at 1 fps and feed them as input images. Following Vidi [14], we design a simple custom prompt with in-context examples, ensuring that the output only contains frame index ranges. An example instruction is shown as

```
The input images are frames from a video. Output the frame indexes that
correspond to the text query: "QUERY". Only output the index range, for
example, 2-4, 6-8.
```

We observe that GPT-5 follows this prompt reliably, typically producing clean and parseable frame ranges, which are then converted into time intervals for evaluation.

We present the detailed results in terms of length, query format, and modality subsets in Table 7. Notably, Vidi2 significantly outperforms Gemini 3 Pro (Preview) and GPT-5 for overall IoU ($> 10\%$). For detailed analysis on different attributes, Vidi2 also excels on IoU, precision, recall on most subsets. We can conclude that a higher overall IoU of Vidi2 score ensures a better experience for automatic video editing/trimming scenarios.

We also present the ROC curve for different metrics in Figure 5, and we can observe that Vidi2 consistently outperforms all competitors in IOU, precision, and recall across the entire range of thresholds. It shows Vidi2's strength in fine-grained temporal understanding, a critical capability for real-world video editing tasks.

4.3 Generic Video QA

As a video understanding foundation model, Vidi2 also supports video question answering, which is a major improvement over the previous Vidi model. We evaluate Vidi2 and several popular multimodal models on three representative video QA benchmarks: LVBench [15], LongVideoBench [17], and VideoMME [4]. While Vidi2 still lags behind the leading proprietary model, Gemini-2.5-Pro [3], in general video QA performance, it achieves competitive results compared to the most popular open-source multimodal model of similar scale, namely Qwen2.5-VL-7B [13]. These results demonstrate that Vidi2 possesses solid multimodal reasoning ability despite being primarily optimized for retrieval and grounding tasks.

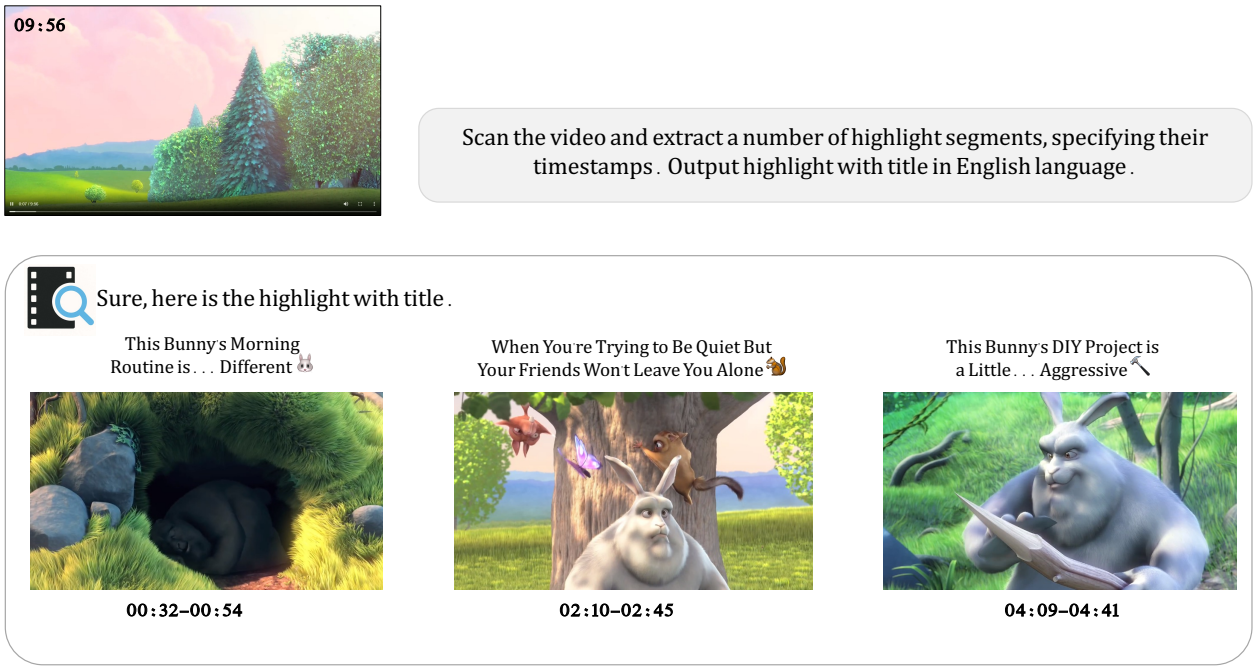


Figure 6: Example of highlight extraction application.

5 Appplications

In view of Vidi2's strong capabilities in temporal retrieval, spatio-temporal grounding, and generic video question answering, it is well suited to support a wide range of video editing tasks, including highlight extraction, plot understanding, and storyline-based video creation.

5.1 Highlight with Title

The strong temporal modeling capability of Vidi2 enables straightforward adaptation to downstream tasks such as highlight prediction. Given a long raw video, the model can automatically generate multiple highlight segments along with concise titles, allowing users to directly publish the resulting short clips. As shown in Figure 6, the generated titles align closely with the corresponding video content, demonstrating both temporal precision and semantic understanding.

5.2 Plot Understanding

Popular movies and TV series often feature complex relationships among multiple characters. Performing character- or plot-based video cutting is therefore time-consuming, even for professional editors. To enable automatic editing in such contexts, a model must comprehend character identity, interactions, and event causality—for example, understanding that "one character raises an arm to hit another but misses and strikes a friend instead." Such scenarios require robust spatio-temporal perception, character recognition, commonsense reasoning, and background knowledge. As shown in Figure 7, Vidi2 supports plot-level understanding and reasoning through a dedicated pipeline, demonstrating its potential in real-world storytelling applications.

5.3 Storyline-based Video Creation

Creating an engaging short video is akin to producing a miniature film with a coherent narrative. A high-quality video depends not only on appealing content but also on an effective storyline and professional-level editing composition. Next-generation video creation systems should act like professional editors, *i.e.*, cutting clips according to a narrative structure while considering package elements such as

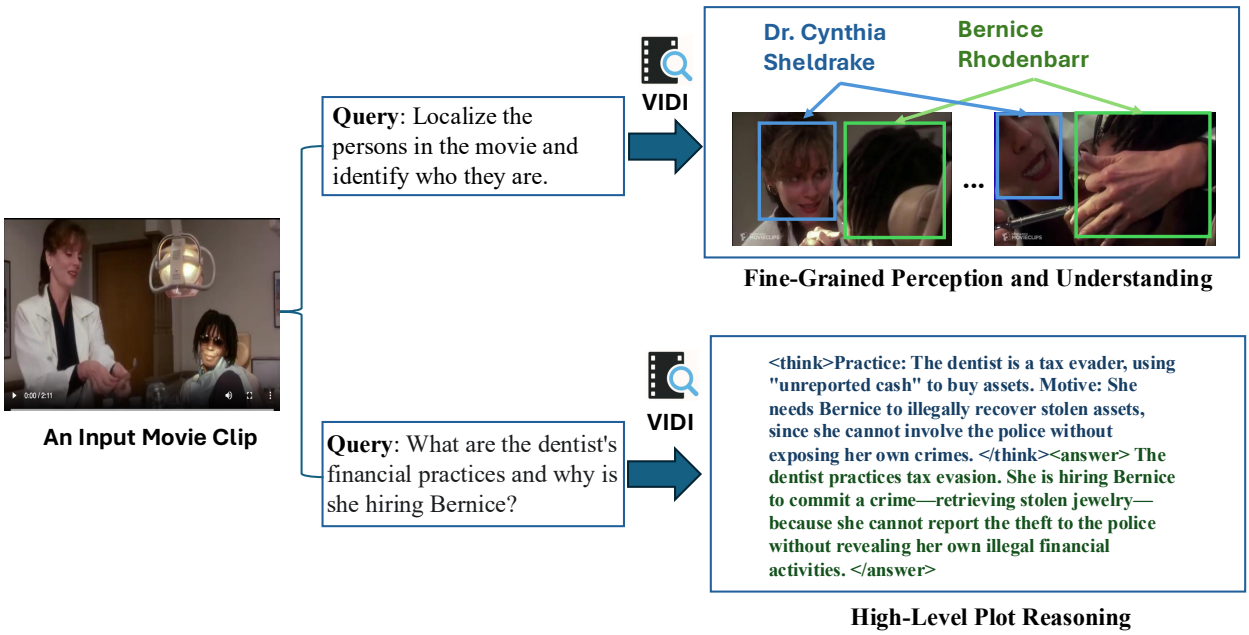
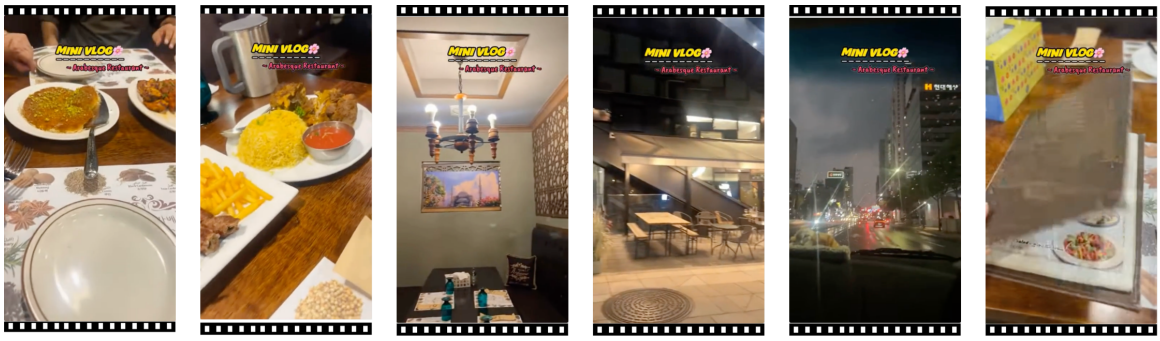


Figure 7: Example of plot understanding application.

effects, transitions, music, and stickers. Vidi2 demonstrates the ability to generate end-to-end video creation instructions conditioned on multiple input videos. As shown in Figure 8, the model takes six videos as input and outputs a complete instruction set for producing a publishable, storyline-driven video. The final rendered output includes narration, music, animations, and transitions, showcasing the model's potential to automate the entire creative editing workflow.

6 Conclusion

In this report, we present Vidi2, which achieves unprecedented end-to-end spatio-temporal grounding performance while maintaining strong video question answering capabilities. It serves as a foundation model for next-generation video creation and understanding scenarios. To evaluate models in realistic video editing contexts, we introduce VUE-STG, a comprehensive benchmark that advances prior spatio-temporal grounding datasets in four aspects: video duration, query format, annotation quality, and evaluation metrics. We also upgrade our previous temporal retrieval benchmark to VUE-TR-V2, featuring a more balanced video-length distribution (including more long and ultra-long videos) and more realistic, user-style queries. In addition, we evaluate Vidi2 on public video QA datasets to assess its general multimodal reasoning ability. Notably, Vidi2 significantly outperforms proprietary systems such as Gemini 3 Pro (Preview) and GPT-5 on both VUE-STG and VUE-TR-V2, demonstrating state-of-the-art spatio-temporal understanding. It also performs on par with the leading open-source video LLM of similar scale, *i.e.*, Qwen2.5-VL-7B, on video QA. Finally, we outline three promising real-world applications that Vidi2 could enable in future demonstrations. Together, these results highlight the model's strong spatio-temporal understanding and multimodal reasoning capabilities, forming a robust foundation for intelligent, human-like video editing and creation systems.



Using the given assets, craft a compelling narrative with a clear emotional arc. Uncover a creative, hidden story instead of just sequencing clips.

Here is the detailed script with narratives to create a high-quality video.

"scenario": "This video is framed as a 'Best Friend's Observation Journal,' humorously documenting his friend's ...
"narrative structure": "The story follows the 'Collection of Moments' blueprint. It opens by establishing the ironic ...
"emotional core": "The video's core is the warm, teasing affection between close friends. The humor comes from ...
"viral strategy": "The relatability of saying you're not hungry but then ordering a feast is a universally shared ...
},
"editing_script":[
{
"shot_timestamp": Asset 3 00:15, 00:19,
"speed":1.0,
"voiceover":"My friend said she 'wasn't even hungry.'",
"keep_sound":false,
"subtitle_style":"Clean sans-serif font, white with a slight drop shadow."
},
.....

Rendering Result

Figure 8: Example of storyline-based video creation application.

7 Contributors

Core Contributors - Research (alphabetical order)
Chia-Wen Kuo, Dawei Du, Fan Chen, Guang Chen, Sijie Zhu, Wen Zhong, Xin Gu, Zhenfang Chen.

Core Contributors - Infrastructure (alphabetical order)
Celong Liu, Haojun Zhao, Tong Jin.

Team Leads
Longyin Wen, Xiaohui Shen.

Contributors (alphabetical order)
Chuang Huang, Haoji Zhang, Lingxi Zhang, Lu Guo, Lusha Li, Qihang Fan, Qingyu Chen, Rachel Deng, Stuart Siew, Weiyan Tao, Zuhua Lin.

References

[1] *https://deepmind.google/models/gemini/.*

[2] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, 2019.

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

[4] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025.

[5] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18330–18339, 2024.

[6] Xin Gu, Yaojie Shen, Chenxi Luo, Tiejian Luo, Yan Huang, Yuewei Lin, Heng Fan, and Libo Zhang. Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. *arXiv preprint arXiv:2502.11168*, 2025.

[7] Yang Jin, Zehuan Yuan, Yadong Mu, et al. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *Advances in Neural Information Processing Systems*, 35:29192–29204, 2022.

[8] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23100–23109, 2023.

[9] OpenAI. GPT-5 System Card. Technical report, OpenAI, August 2025. Version published August 7 2025. Available at: `https://openai.com/index/gpt-5-system-card/`.

[10] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542, 2021.

[11] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021.

[12] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[13] Qwen Team. Qwen2.5-vl, January 2025.

[14] Vidi Team, Celong Liu, Chia-Wen Kuo, Dawei Du, Fan Chen, Guang Chen, Jiamin Yuan, Lingxi Zhang, Lu Guo, Lusha Li, et al. Vidi: Large multimodal models for video understanding and editing. *arXiv preprint arXiv:2504.15681*, 2025.

[15] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025.

[16] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18909–18918, 2024.

[17] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[18] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *Proceedings of the IEEE international conference on computer vision*, pages 1453–1462, 2017.

[19] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020.