NJUST-KMG at MedGenVidQA 2026: Cascade Multi-modal Alignment with Gaussian Soft Priors for Medical Visual Answer Localization

Jinglong Li and Yang Yang
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, Jiangsu, China
{jinglong_555, yyang}@njust.edu.cn

Abstract

This paper describes the system developed for the Medical Visual Answer Localization (MVAL) task at MedGenVidQA 2026. Accurately locating surgical or instructional steps in medical videos is inherently challenging due to audio-visual asynchrony and the visual homogeneity of surgical scenes. We propose a Cascade Multi-modal Alignment Framework that integrates Large Language Models (LLMs) to bridge the semantic-temporal gap. Our pipeline utilizes WhisperX for word-level speech transcription to ensure precise textual anchoring. We then employ Gemini3 as a high-level semantic ranker to generate multi-scale textual priors. Crucially, we transform these discrete semantic scores into a continuous 1D Gaussian Soft Prior, which is injected as an attention bias into our cross-modal fusion network. This mechanism preserves global temporal context while guiding the model to focus on query-relevant frames. Our system achieves highly competitive performance on the validation leaderboard, particularly under strict evaluation metrics, reaching an IoU@0.7 of 67.5%.

1 Introduction

Temporal Video Grounding (TVG) in the medical domain, specifically Visual Answer Localization (VAL), requires a system to predict the precise start and end timestamps of a video segment that answers a given medical query. Unlike generic instructional videos, medical and surgical videos often feature high visual redundancy and "semantic drift," where the practitioner's verbal explanation does not perfectly align with the visual execution of a step.

Most existing cascade localization systems rely on "hard truncation," where a text-retrieval module first crops the video into short candidate clips. While efficient, this approach is prone to "boundary collapse" if the initial text-based window is too narrow or slightly shifted, leading to significant performance degradation at high Intersection over Union (IoU) thresholds.

To overcome these limitations, we propose a soft-prior alignment strategy inspired by recent successes in event segmentation(Zhai et al., 2020). Our contributions are three-fold:

1. We implement a precision ASR pipeline using WhisperX to extract word-level timestamps, ensuring that the foundational textual data is accurately anchored to the video timeline.

2. We design a semantic scoring module using Gemini3 that aggregates multi-scale textual windows to capture both local actions and global surgical stages.

3. We introduce a 1D Gaussian Soft Prior that serves as an attention guidance mechanism, allowing the visual module to refine boundaries beyond the speech-indicated segments.

2 Related Work

2.1 Medical Visual Answer Localization

The MedVidQA and MedGenVidQA benchmarks(Gupta and Demner-Fushman, 2022; Gupta et al., 2023, 2025, 2026) have established the Medical Visual Answer Localization (MVAL) task as a core challenge in multi-modal healthcare AI. Early approaches relied heavily on pure visual span predictors(Gao et al., 2017; Zhang et al., 2020), which often struggled with the visual homogeneity of endoscopic or laparoscopic footage. Subsequent research demonstrated that leveraging ASR transcripts can act as a powerful prior for localization.

2.2 LLMs for Multi-modal Grounding

Recent trends in TRECVID challenges show a shift toward using Large Language Models (LLMs) as zero-shot or few-shot reasoners. Specifically, the use of LLMs to generate temporal "proposals"

Proceedings of the BioNLP 2026 (Shared Tasks), pages 229–232
July 3-4, 2026 ©2026 Association for Computational Linguistics

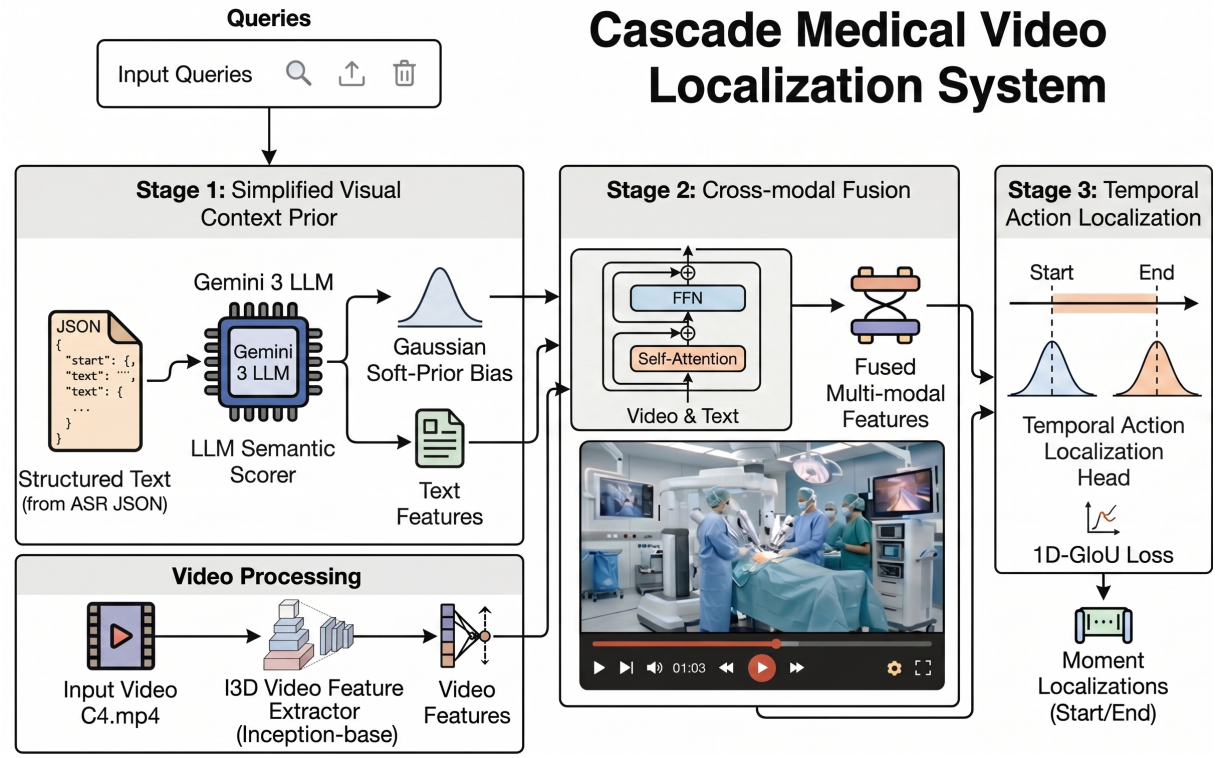

Figure 1: Overall architecture of the proposed Cascade Medical Video Localization System. The framework leverages a 1D Gaussian Soft Prior to inject textual relevance directly into the cross-modal fusion transformer via Attention Bias, ensuring high-precision boundary regression without the risks of hard truncation.

or "summaries" has proven effective(Zhang et al., 2023; Ren et al., 2023). Our work builds upon these insights by moving from discrete proposal generation to continuous probabilistic weighting.

3 Methodology

Figure 1 illustrates the complete architecture of our proposed Cascade Multi-modal Alignment Framework. It consists of three main stages: (1) Precision ASR extraction and multi-scale aggregation; (2) LLM-based semantic scoring and 1D Gaussian Soft Prior generation; and (3) Cross-modal fusion with attention biasing for final localization.

3.1 ASR and Multi-scale Construction

As shown in Figure 2, the frontend utilizes a specialized pipeline for audio-to-text transcription. We employ WhisperX, which improves upon standard Whisper models by using forced alignment (e.g., via Wav2Vec2) to provide word-level timestamps.

For each identified segment s_i with text T_i, we define the interval as $[t^i_{start}, t^i_{end}]$. To capture varying semantic granularities and mitigate the issue of fragmented speech, we construct multi-scale windows $W_{m,k}$ by aggregating m consecutive segments. This ensures that the system can capture quick actions (e.g., "cutting a suture") as well as prolonged stages (e.g., "organ dissection").

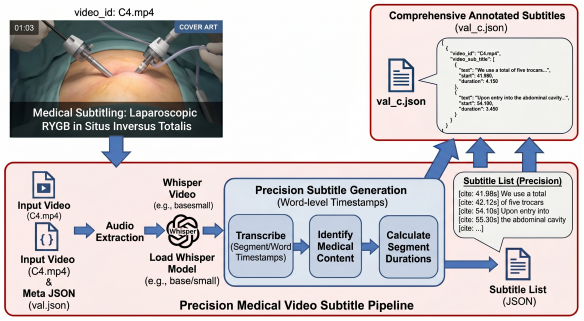

Figure 2: The Precision Medical Video Subtitle Pipeline utilizing WhisperX/Wav2Vec2 for extracting forced word-level timestamps.

3.2 LLM Semantic Ranking

We utilize Gemini3 to calculate the relevance score S_j for each window W_j relative to the medical query Q. Instead of direct generation, we extract the logits for the positive ("Yes") token to obtain a continuous probability $p_j = P(\text{Relevant}|Q, W_j)$. This provides a more granular signal than binary classification.

3.3 1D Gaussian Soft Prior Modeling

Given the highest-scoring window $[t_{start}, t_{end}]$, we define the temporal center $\mu = (t_{start} + t_{end})/2$. To allow the model to explore boundaries outside this window, we model the temporal prior as a 1D Gaussian distribution $G(t)$:

$$G(t) = \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \quad (1)$$

where σ is a hyperparameter that controls the "width" of the soft prior. This distribution represents the probability that a specific frame at time t belongs to the answer segment according to the linguistic evidence.

3.4 Cross-modal Attention Biasing

We extract dense spatio-temporal visual features $V \in \mathbb{R}^{L \times d}$ using a pre-trained I3D (Inflated 3D ConvNet) encoder. Unlike frame-level spatial extractors, I3D inherently captures local motion dynamics, which synergizes with our temporal soft prior. Let H_v be the visual hidden states and H_q be the query embedding. In the cross-modal fusion layer (Stage 2 in Figure 1), we inject $G(t)$ as an **Attention Bias**. Modifying the pre-softmax attention scores with explicit structural biases has been shown to effectively guide the attention distribution (Press et al., 2022). The standard cross-attention matrix A is modified as follows:

$$A_{i,j} = \text{Softmax}\left(\frac{Q_i K_j^T}{\sqrt{d}} + \lambda \cdot \log(G_j + \epsilon)\right) \quad (2)$$

where λ is a scaling factor. This bias guides the attention mechanism to prioritize frames within the high-confidence Gaussian region while still permitting the visual network to attend to outlying frames if the visual evidence is strong.

4 Experimental Setup

4.1 Datasets

We evaluate our system on the MedGenVidQA 2026 validation set. Evaluation is based on IoU at thresholds $\{0.3, 0.5, 0.7\}$ and mean IoU (mIoU).

4.2 Implementation Details

For the visual stream, the I3D feature extractor processes the video using a 16-frame sliding window without overlap, yielding a continuous feature sequence that perfectly aligns with our 1D Gaussian bias temporal resolution. We utilize LoRA(Hu et al., 2021) with a rank $r = 32$ for parameter-efficient adaptation. Detailed hyperparameters are provided in Table 1.

Hyperparameter	Value
Visual Backbone	I3D
LoRA Rank (r)	32
LoRA Alpha (α)	64
Optimizer	AdamW
Learning Rate	1×10^{-4}
Gaussian σ	Dynamic
Bias Scale λ	0.5

Table 1: Hyperparameter configurations for the proposed system.

5 Results and Analysis

5.1 Quantitative Results

Table 2 compares our "Soft Prior" approach with baseline methods. Our system significantly outperforms the "Hard Truncation" baseline, especially at the strict IoU@0.7 threshold.

Approach	IoU@0.3	IoU@0.5	IoU@0.7
Pure Visual	68.75	52.5	26.25
Hard Truncation	86.25	75.0	60
Soft Prior	**92.5**	**81.25**	**67.5**

Table 2: Comparative performance results on MedGen-VidQA.

5.2 The Impact of Soft Priors

The primary advantage of the 1D Gaussian Soft Prior is its ability to handle asynchrony. In surgical instructional videos, a surgeon might say "Now I am suturing the incision" while the visual action of suturing continues for several seconds after the speech ends. Hard truncation would lose the end of this action, whereas our soft prior allows the visual module to "track" the action to its true conclusion.

5.3 Error Analysis

We observed performance drops in two specific scenarios: (1) **Silent Videos**: In videos where instructions are purely visual or provided via on-screen text, our ASR-based prior fails. (2) **Implicit Semantics**: In cases involving sarcasm or complex medical metaphors, the LLM scoring becomes noisy, leading to a flattened Gaussian distribution.

6 Discussion and Future Work

Our findings align with previous work in TRECVID QFISC(Awad et al., 2023), suggesting that the audio modality is often more informative for step boundary detection than the raw visual stream. However, the fusion of these modalities remains non-trivial. Future work will explore Mixture-of-Experts (MoE) architectures to dynamically weight the audio and visual streams based on the detected presence of speech or on-screen text.

7 Conclusion

Team NJUST-KMG's submission for MedGen-VidQA 2026 leverages a cascade multi-modal framework. By combining precision ASR extraction with LLM-driven Gaussian soft priors, we successfully bridge the semantic gap in medical video localization. Our results demonstrate that replacing rigid temporal cropping with attention-guided soft priors is a superior strategy for achieving high-precision boundary regression.

References

George Awad, Asad A Butt, Jonathan Fiscus, Martial Michel, David Joy, and 1 others. 2023. Trecvid 2023: Evaluation of video activity detection, video captioning and retrieval, and video to text. *arXiv preprint arXiv:2401.xxxx*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025. A dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1):1035.

Deepak Gupta, Collin Scott Campbell, Pedram Golnari, and Dina Demner-Fushman. 2026. Overview of the medgenvidqa 2026 shared task on medical generative video question answering. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, San Diego, USA. Association for Computational Linguistics.

Deepak Gupta and Dina Demner-Fushman. 2022. Medvidqa: A multilingual and multimodal dataset for medical video question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1302–1318.

Edward J Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ofir Press, Noah A Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Shuhuai Ren and 1 others. 2023. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*.

Shao Huang Zhai and 1 others. 2020. Event segmentation with action boundaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1011–1020.

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

Songyang Zhang, Houwen Peng, Jianlong Jian, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877.