

VSI: Visual-Subtitle Integration for Keyframe Selection to Enhance Long Video Understanding

To Reviewer QREu08:

Loose Keyframe Evaluation Criterion To address this, we adopted the stricter **SSIM Precision**, **SSIM Recall** and **SSIM F1 Score** metrics to evaluate visual content fidelity on the full LongVideoBench dataset. As shown in the table below, VSI consistently outperforms SOTA baselines.

Metric	Uni-8	Tstar-8	VSLs-8	VSI-8	Uni-32	Tstar-32	VSLs-32	VSI-32
SSIM Precision \uparrow	60.7	75.3	75.6	76.8	60.2	74.0	74.5	75.2
SSIM Recall \uparrow	80.4	88.2	88.6	89.7	85.0	90.3	92.5	94.3
SSIM F1 Score \uparrow	69.2	81.2	81.6	82.8	70.5	81.3	82.5	83.7

Clarity, Notation, and Organization We have revised the manuscript to address these points: (1) **Figure 1**: Updated the caption to label it as a schematic illustration explicitly; (2) **Notation**: Disambiguated symbols by renaming subtitles to $\mathcal{C} = \{c_i\}$, object sets to \mathcal{O}_t , similarity threshold to τ_{text} , and computational budget to B_{max} ; (3) **Eq. (15)**: Clarified that Eq. (15) computes sampling probability for exploration, while final selection relies on accumulated confidence scores (S_{fused}) for exploitation.

To Reviewer Ewk312 & BXoY08 & QREu08:

Subtitle Dependency and Robustness The dual-branch architecture is not a simple summation but a collaborative mechanism where the Subtitle branch provides a coarse temporal prior to guide the Visual branch towards semantically relevant windows, while the Visual branch performs fine-grained frame localization. We demonstrate VSI’s robustness through both theoretical guarantees and empirical stress tests. As shown in the table below, in the **No Subtitle** setting, due to the lack of guidance from the Subtitle branch, Recall experienced a significant decline, but Precision remained at its original level, indicating that the predicted frames still effectively cover the visual features of the Ground Truth frames. In the **Noisy Subtitle** setting, the introduction of false semantic cues causes a slight distraction and lowers both Precision and Recall, yet the system does not collapse. This confirms that VSI serves as a safe enhancement: it significantly boosts performance when high-quality text is available while ensuring robust fallback capability in silent or noisy scenarios.

Setting	Subtitle Availability	Frames	Precision \uparrow	Recall \uparrow	F1 Score \uparrow
VSI (Standard)	100%	32	75.2	94.3	83.7
VSI (No Subtitle)	0%	32	75.0	93.5	83.2
VSI (Noisy Subtitle)	Random	32	74.6	92.7	82.7
Baseline (VSLs)	N/A	32	74.5	92.5	82.5
Baseline (Tstar)	N/A	32	74.0	90.3	81.3

Effectiveness of Video Search Branch and Fusion Sensitivity The performance drop in text-referred tasks (Tables 3 and 5) is expected, as these tasks (T2A, T2E, T2O) rely predominantly on dialogue timestamps, making additional visual search a source of modality competition in these specific subsets. However, on the full dataset, the

Video Search branch proves indispensable. As shown in the table below, the Subtitle Match branch excels in Recall (94.1), identifying candidate temporal windows, while the Video Search branch achieves superior Precision (75.0), effectively pinpointing the visually correct frames. The Fusion model achieves the highest F1 Score (83.7), confirming that the Video Search branch provides essential visual precision rather than noise, complementing the subtitle branch to maximize global effectiveness.

Setting	Frames	Precision \uparrow	Recall \uparrow	F1 Score \uparrow
VSI (Fusion)	32	75.2	94.3	83.7
VSI (Subtitle Match)	32	74.3	94.1	83.0
VSI (Video Search)	32	75.0	93.5	83.2

To Reviewer Ewk312 & BXoY08:

Computational Cost and Latency The computational We address efficiency concerns by highlighting that, despite the dual-branch design, VSI (31.71s) is actually faster than the previous SOTA baseline VSLs (33.26s) on the full dataset while achieving superior accuracy (73.89%). This efficiency stems from the fact that the Subtitle Match branch and cross-modal fusion introduce negligible $O(1)$ overhead, with the computational bottleneck remaining the visual backbone shared by existing methods.

To Reviewer BXoY08:

Redundancy Analysis and Branch Overlap VSI fuses scores from both branches to generate a unified sampling probability distribution, rather than maintaining separate candidate lists. Consequently, overlap simply means a frame scores high in both modalities; such frames naturally gain higher sampling probabilities as they capture both visual and semantic relevance. In cases of conflict, the final score is determined by the fusion weight, which balances the focus between visual and textual signals according to the task type. We calculated the average overlap rate of the top-K key frames selected by the two branches, which was only **34.60%**. This high degree of complementarity enables VSI to cover the 65.40% of key information points that would inevitably be missed by a single modality.

Baselines for Retrieval : We expanded the evaluation to include Dense Retrieval (VideoAgent[2]) and Temporal Searching (AKS[1]) methods. Under the 8-frame setting, VSI achieves the highest SSIM F1 Score of **82.8**, significantly outperforming VideoAgent (65.2) and AKS (80.4).

References

- [1] Tang X et al. Adaptive keyframe sampling for long video understanding, 2025. 1
- [2] Wang XH et al. Videoagent: Long-form video understanding with large language model as agent, 2024. 1