Follow the Saliency: Supervised Saliency for Retrieval-augmented Dense Video Captioning

Supplementary Material

In this Appendix, we present additional experiments, detailed formulations, and qualitative results to further support our findings on the proposed *STaRC*. Specifically, §A provides additional analyses and experiments on *Supervised Saliency Training*, §B presents extended formulations and studies for *Saliency-Guided Segmentation and Retrieval (SGSR)*, and §C additional qualitative results and visualizations, particularly focusing on experiments regarding *Sliding Window Self-Attention (SWSA)* module.

A. Supervised Saliency Training

§A.1 analysis the impact of hyperparameters (λ, τ) in the saliency loss and §A.2 offers visualizations of predicted saliency scores.

A.1. Saliency Loss Hyperparameters

The hyperparameters λ and τ significantly impact saliency learning performance, as shown in Figure A.1. These hyperparameters appear in our saliency loss formulation (**??** and **??**), where τ controls the softmax temperature and λ weights the saliency loss. YouCook2 [8] achieves best performance at $\lambda = 6.0$, while ViTT [1] performs best at $\lambda = 2.0$. For the temperature parameter τ, 0.5 provides optimal performance.

A.2. Qualitative Analysis of Saliency Scores

We visualize the predicted saliency scores from the trained highlight detection module in Figure A.2. Saliency scores are high within ground truth event boundaries and low in non-event regions. Quantitative analysis confirms that event regions show significantly higher mean scores than non-event regions, demonstrating effective separation of important frames from background.

B. Saliency-Guided Segmentation and Retrieval

§B.1 provides detailed equations for the saliency-aware OT-based segmentation method in *SGSR*, including Kantorovich OT matching, Gromov-Wasserstein OT, fused OT objective, and the balanced-unbalanced constraint that uniquely differentiates *SGSR* from prior OT-based formulations such as ASOT [7], along with hyperparameter experiments. §B.2 presents the saliency weighted pooling equation and qualitative segmentation results with retrieved captions.

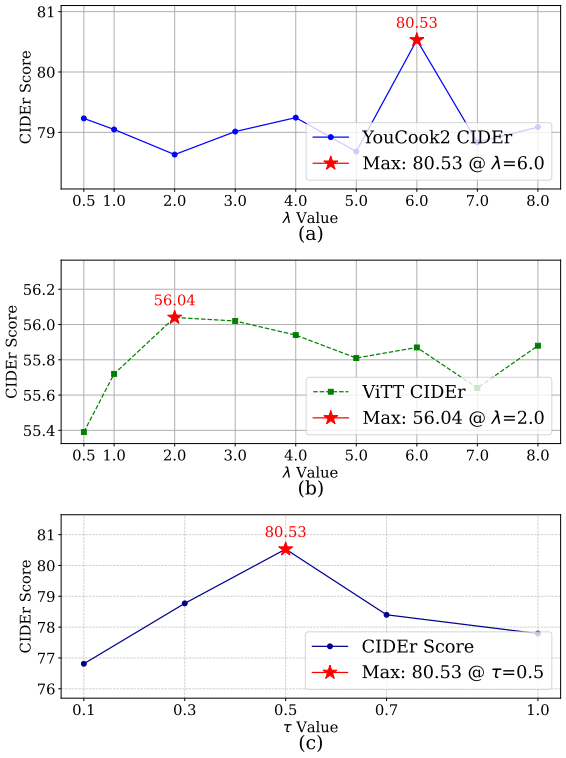

Figure A.1. **(a-b) CIDEr performance for different saliency-loss weights λ.** Results on YouCook2 and ViTT showing how different values of λ influence the strength of supervised saliency learning and overall caption quality. **(c) CIDEr performance in YouCook2 validation set for different temperature values τ.** This figure visualizes model sensitivity to different τ values used in the loss formulation.

Preliminary. Given frame spatial features $X^s \in \mathbb{R}^{F\times D}$ and learnable anchors $A = \{a_j\}_{j=1}^{K} \in \mathbb{R}^{K\times D}$, we solve for a soft assignment matrix $T \in \mathbb{F}^{N\times K}$, where T_{ij} denotes the probability that frame i belongs to prototype j.

B.1. OT-based segmentation method

Saliency-aware Kantorovich OT matching. We formulate the OT problem using the Kantorovich objective [6]. Let the anchor and frame marginals be defined as $\mathbf{p} = \frac{1}{F}1_F$ and $\mathbf{q} = \frac{1}{K}1_K$:

$$\underset{\mathbf{T}\in\mathcal{T}_{\mathbf{p},\mathbf{q}}}{\text{minimize}} \quad \mathcal{F}_{\text{KOT}}(\mathbf{C}^k, \mathbf{T}) := \langle \mathbf{C}^k, \mathbf{T}\rangle. \tag{1}$$

The visual affinity between frames and anchors is measured using cosine similarity. To incorporate saliency infor-

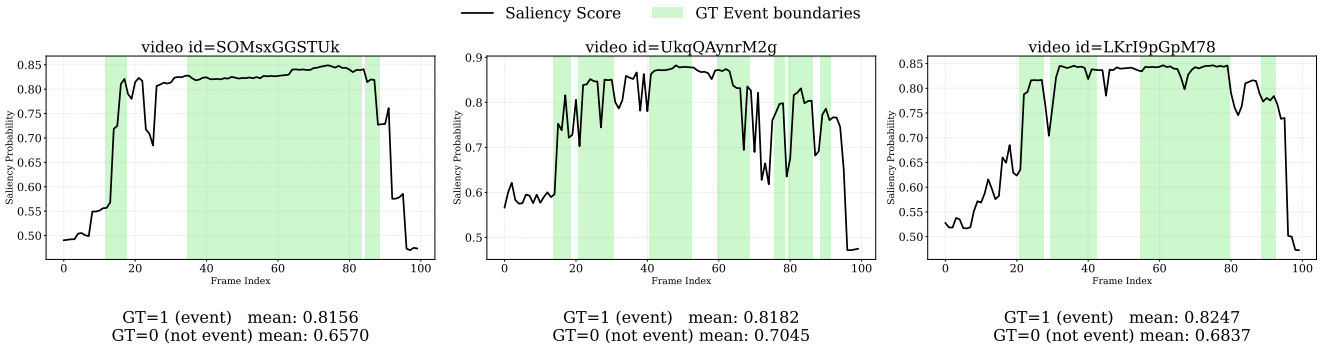

Figure A.2. **Visualization of predicted frame-level saliency scores.** Black lines show predicted saliency scores overlaid with ground truth event boundaries (green regions). Saliency scores are higher within event regions than in non-event regions. This separation confirms that the model successfully learns to align saliency prediction with true event boundaries.

mation, we introduce a bias term weighted by μ using the saliency prior p_s. This yields the following cost matrix:

$$C^k_{nj} := (1 - \frac{{x^s_n}^\top a_j}{\|x^s_n\|_2 \|a_j\|_2}) - \mu p_{s_n}. \quad (2)$$

Temporal consistency via Gromov-Wasserstein OT. To enforce smooth temporal structure, we adopt the Gromov-Wasserstein (GW) OT [5]. Let C^v capture temporal proximity between frames and C^a captures the structural relationship between prototypes. We define $[F] := \{1, ..., F\}$. The GW objective is then formulated as:

$$\mathcal{F}_{\text{GW}}(\mathbf{C}^v, \mathbf{C}^a, \mathbf{T}) := \sum_{\substack{n,F\in[F]\\ j,K\in A}} L(C^v_{nF}, C^a_{jK}) T_{nj} T_{FK}. \quad (3)$$

Fused GW Optimal Transport Objective. The final *SGSR* objective combines visual matching and structural consistency as Fused GW OT objective:

$$\mathcal{F}_{\text{FGW}}(\mathbf{C}, \mathbf{T}) := \alpha \mathcal{F}_{\text{GW}}(\mathbf{C}^v, \mathbf{C}^a, \mathbf{T}) + (1-\alpha)\mathcal{F}_{\text{KOT}}(\mathbf{C}^k, \mathbf{T}). \quad (4)$$

Unbalanced OT. Unlike ASOT, we reverse the balanced/unbalanced configuration to better suit the DVC task. Specifically, we apply unbalanced OT for frame-level matching using saliency scores as the marginal constraint $\mathbf{p}$, while the anchor marginal remains balanced, as formulated in Equation (5):

$$\min_{\mathbf{T}\in\mathcal{T}_q} \mathcal{F}_{\text{FGW}}(\mathbf{C}, \mathbf{T}) + \gamma \, \mathrm{D}_{\text{KL}}\big(\mathbf{T}^\top \mathbf{1}_K \,\big\|\, p_s\big). \quad (5)$$

This configuration enables the transport mass to follow the saliency distribution, allowing the model to adaptively allocate attention to salient regions. As shown in Figure A.3, this reversed configuration achieves the best performance for our task. Therefore, our *SGSR* setting of applying unbalanced OT with saliency score constraints for frames and balanced OT for anchors proves to be optimal for DVC.

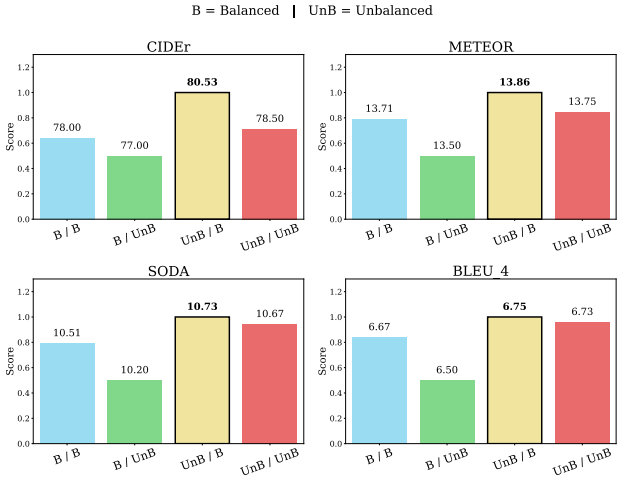

Figure A.3. **Performance comparison of *SGSR* under different OT configurations.** We evaluate Balanced (B) and Unbalanced (UnB) OT for both frame-side and anchor-side matching. The configuration with UnB for frames and B for anchors achieves the best performance across all metrics.

Hyperparameter Analysis. Except for the saliency-related terms μ and γ, we keep all other OT parameters identical to ASOT. We set the unbalanced regularization coefficient ρ to 0 during both training and inference. This improves CIDEr from 79.61 to 80.53. We analyze different μ and γ combinations in Figure A.4. The best performance is achieved at $\mu = 0.1$ and $\gamma = 0.3$. This shows that moderate

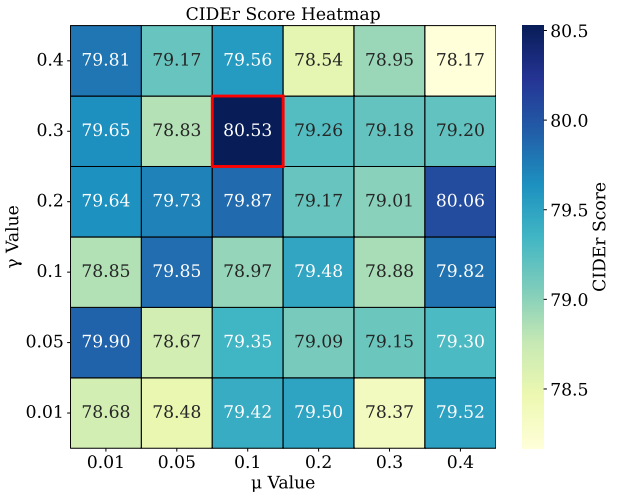

Figure A.4. **CIDEr performance heatmap across different combinations of saliency weight μ and temporal-consistency weight γ.** Each cell shows the CIDEr score obtained under a specific μ and γ configuration in the *SGSR* objective.

Datstore Type	YouCook2 (val)			
	CIDEr	METEOR	SODA_c	BLEU_4
CC3M	76.65	13.63	10.53	6.63
COCO	77.31	13.57	10.66	6.67
Hierarchical [4]	77.28	13.59	10.42	6.59
In-domain	**80.53**	**13.86**	**10.73**	**6.75**

Table A.1. Ablation study on different datstores used for retrieval.

saliency weighting and temporal consistency work best for our task.

B.2. Retrieval process

Representative Segment Features for Retrieval. After obtaining the optimized transport plan T^*, we generate segments. We rank all segments using the scoring and select the top-k segments for retrieval. For each selected segment, we construct a representative feature by aggregating its frame embeddings using a saliency-weighted average to emphasize informative frames.

Let $s_j = [f_j^s, f_j^e)$ denote the frame indices assigned to segment j. The representative feature is computed as:

$$\bar{x}_j^s = \frac{\sum_{n \in s_j} p_{s,n}\, x_n^s}{\sum_{n \in s_j} p_{s,n}}, \tag{6}$$

These k representative features are then used for retrieval following the mechanism described in **??**. As shown in **??**, the saliency-weighted pooling leads to more accurate segment representations for retrieval. This results in consistent improvements in overall DVC performance compared to the uniform averaging strategy used in previous work [2–4].

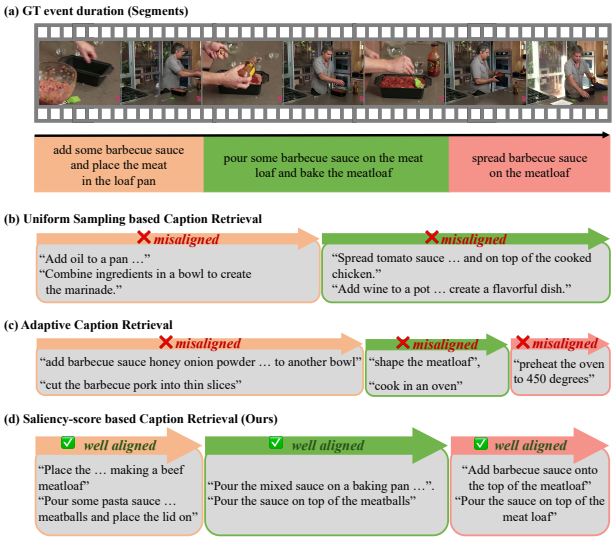

Figure A.5. **Qualitative comparison of retrieved captions under different segmentation strategies.** (a) Ground truth segments and captions. (b) Segments from uniform sampling (HiCM2 [4]) and their retrieved captions. (c) Segments from adaptive sampling (Sali4Vid [2]) and their retrieved captions. (d) Our *SGSR* produces segments that align more closely with true event boundaries and retrieves more relevant captions.

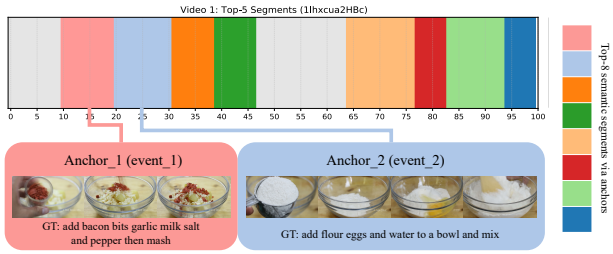

Figure A.6. **Visualization of the Semantic Prototypes.**

In addition, we evaluate retrieval performance under different datastores to assess the robustness of our method. As reported in Table A.1, *SGSR* consistently maintains state-of-the-art performance.

Segment and Retrieved Captions Qualitative Results. We qualitatively compare segments and retrieved captions in Figure A.5. While prior methods [2, 4] often produce misaligned or drifting segment boundaries, our saliency-guided transport yields segments that align more closely with ground truth event durations. The retrieved captions are also semantically more relevant to the video content, validating the effectiveness of *SGSR*. In *SGSR*, we define K learnable anchor embeddings as semantic prototypes, where each anchor serves as a general temporal placeholder representing an individual event. During optimal transport,

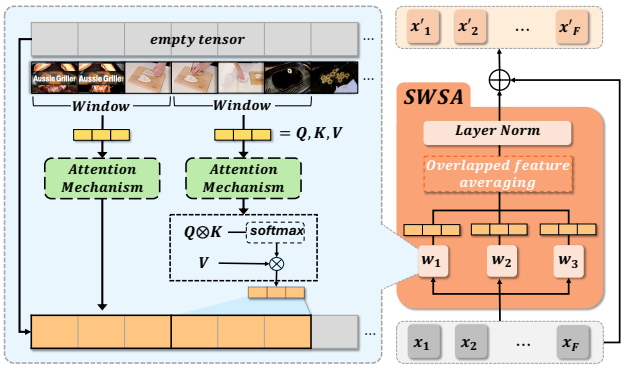

Figure A.7. **Architecture of the *SWSA* module.** *SWSA* refines X through local self-attention without linear projections over sliding windows, where overlapping regions are averaged and residually added to produce refined features X'.

frames are assigned to the anchor with the highest feature similarity, forming temporally coherent segments. We visualize the learned anchor assignments in Figure A.6, showing that the anchors consistently capture meaningful event boundaries.

C. Additional Analysis and Experiments

Feature Refinement. The architecture of *SWSA* is on the Figure A.7.

Given the input features $X \in \mathbb{R}^{F \times D}$ and a set of window sizes $\{w_1, w_2, w_3\}$, we compute local self-attention within each window. For a window of size w, the segment starting at position i is denoted as $X_{i:i+w} \in \mathbb{R}^{w \times D}$. The attention output for this segment is computed as:

$$A_{i:i+w} = \text{softmax}\left(\frac{X_{i:i+w} X_{i:i+w}^{\top}}{\sqrt{D}}\right) X_{i:i+w}. \quad (7)$$

Since windows overlap, each frame n may appear in multiple windows. We average all attention outputs that cover frame n, weighted by the number of overlapping windows C_n:

$$\hat{X}_n = \frac{1}{C_n} \sum_{(w,i):\, n \in [i, i+w)} A_n^{(w,i)}, \quad (8)$$

where C_n is the number of windows covering frame n, and $A_n^{(w,i)}$ is the attention output at position n from the window of size w starting at i. The final refined representation is obtained via a residual connection:

$$X' = X + \text{LayerNorm}(\hat{X}). \quad (9)$$

Feature Difference between Training and Inference. In *STaRC*, we use refined features from *SWSA* differently during training and inference. During training, the refined features X' are used only as input to the highlight detection module, while the decoder receives original features X. During inference, the refined features X' are used both as input to the highlight detection module and as decoder input. To analyze the refinement impact, we measure frame-to-frame transition sharpness in Figure A.8, by computing L2 distances between consecutive frame embeddings. Refined features X' produce sharper transitions at event boundaries compared to X, while non-event regions show smaller transitions.

Method	PT	CIDEr	SODA_c	F1
TimeChat CVPR24	✓	11.0	3.4	19.5
VTG-LLM AAAI25	✓	13.4	3.6	20.6
TRACE ICLR25	✓	35.5	6.7	31.8
TimeExpert ICCV25	✓	39.0	7.2	33.5
Ours	✓	**80.53**	**10.73**	**34.34**

Table A.2. Comparison with VLM methods on the YouCook2 validation set

Comparison with Vision-Language Models. We also compare with recent Vision-Language Models (VLMs) on the DVC task. As shown in Table A.2, even powerful VLMs struggle to produce accurate event boundaries and captions simultaneously. This suggests that general-purpose VLMs are not yet sufficient for DVC, and task-specific architectures remain essential.

D. Failure case and limitations

STaRC occasionally produces redundant captions despite precise localization (a challenge shared by existing DVC methods). OT-based retrieval is effective overall but can introduce noise in some segments (Figure A.9), which we leave for future work.

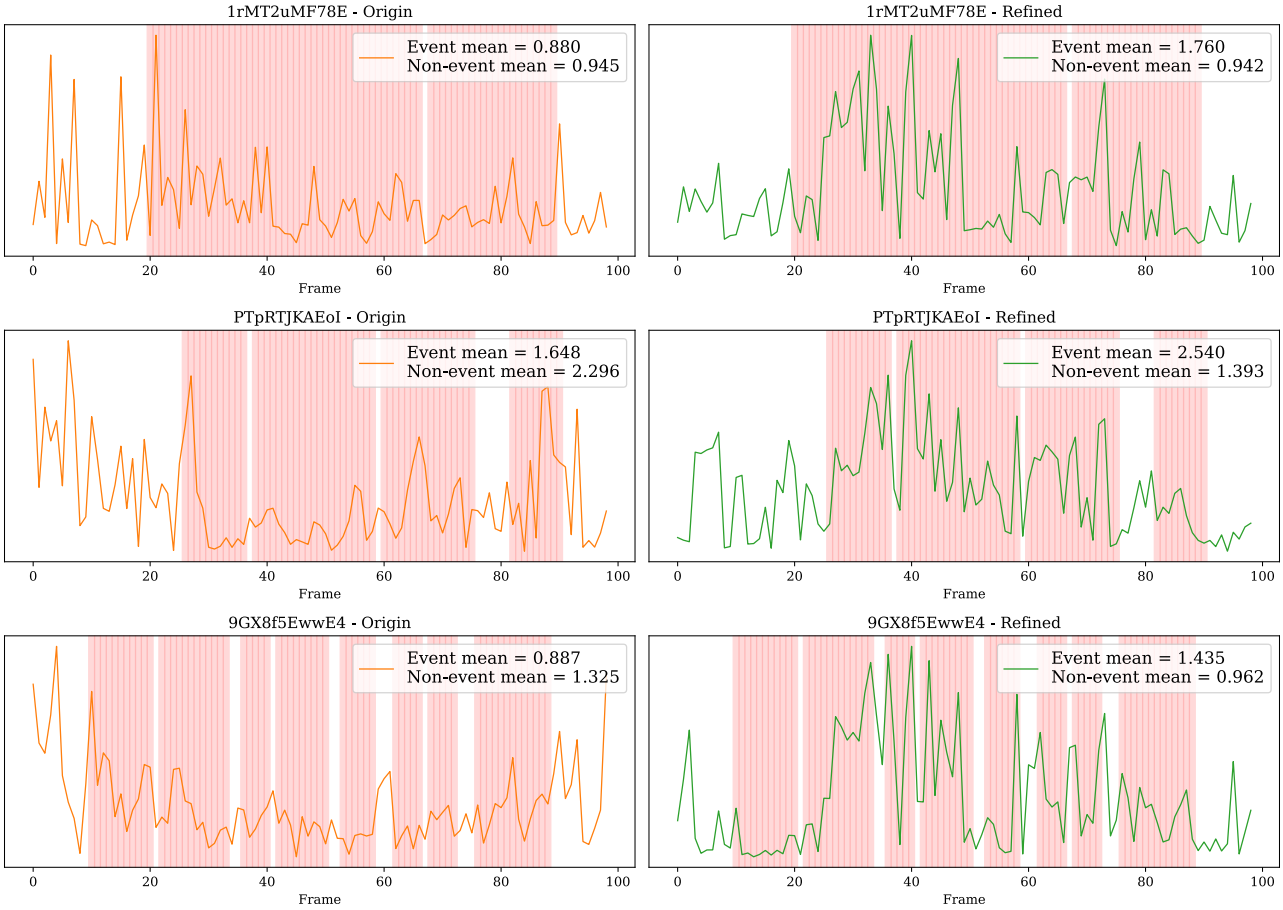

Figure A.8. **Frame transition sharpness comparison between original and refined features**. For each video in the YouCook2 validation set, red spans indicate ground truth event regions. Refined features exhibit larger transition magnitudes within event intervals, indicating stronger semantic contrast and clearer event boundaries than original features.

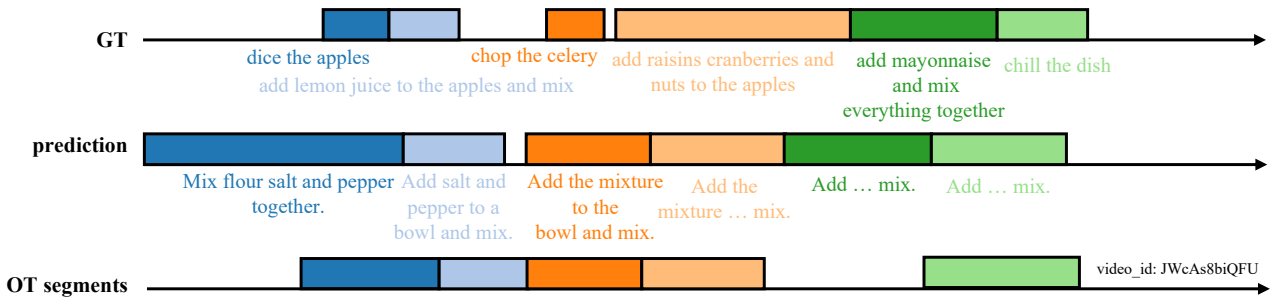

Figure A.9. **Failure case of *STaRC***.

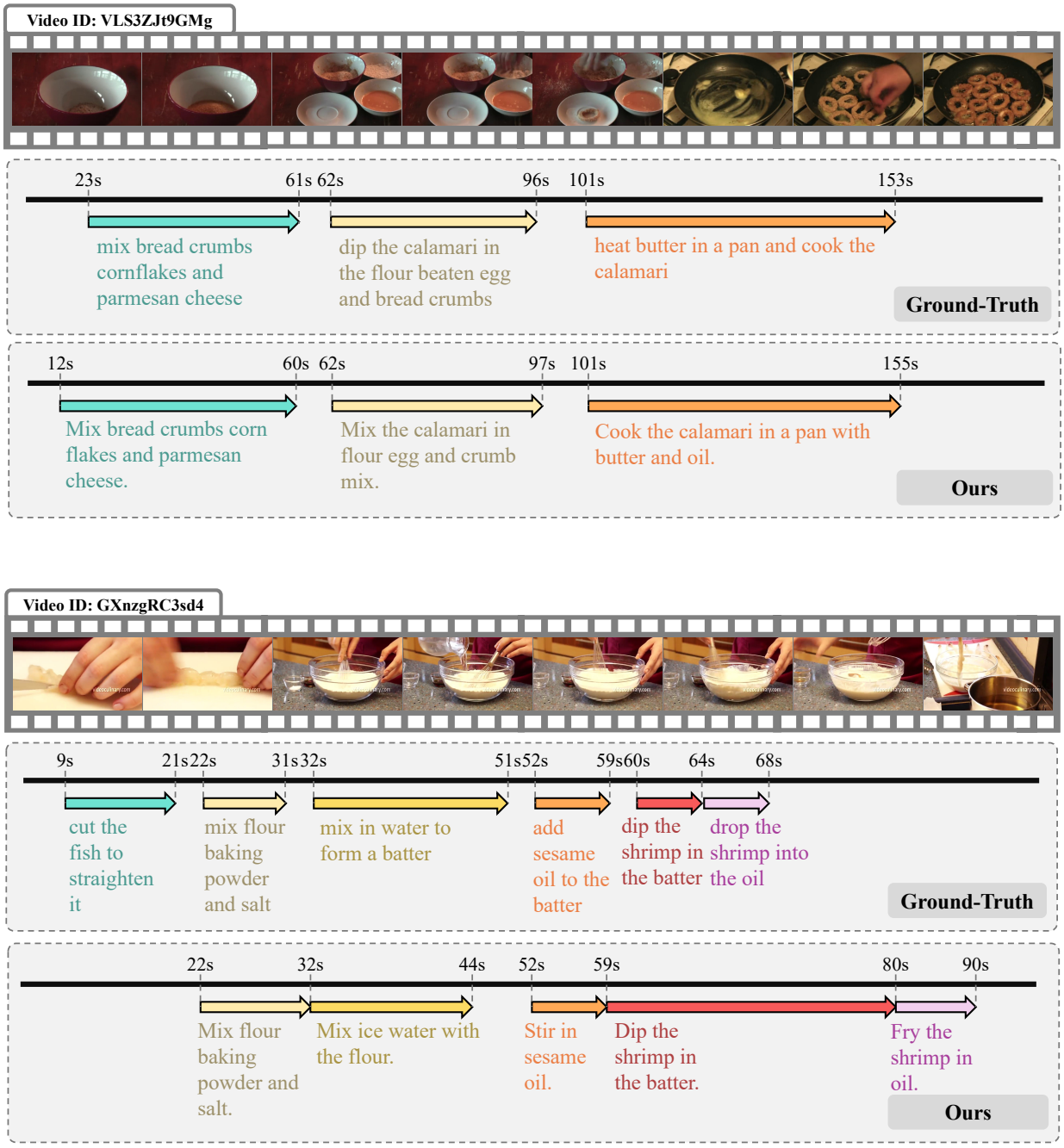

Figure A.10. **Additional qualitative results for *STaRC* on the YouCook2 validation set.**

References

[1] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*, 2020. 1

[2] MinJu Jeon, Si-Woo Kim, Ye-Chan Kim, HyunGee Kim, and Dong-Jin Kim. Sali4vid: Saliency-aware video reweighting and adaptive caption retrieval for dense video captioning. *arXiv preprint arXiv:2509.04602*, 2025. 3

[3] Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13904, 2024.

[4] Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Hicm^2: Hierarchical compact memory modeling for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4293–4301, 2025. 3

[5] Gabriel Peyr, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. 2016. 2

[6] Matthew Thorpe. Introduction to optimal transport, 2018. 1

[7] Ming Xu and Stephen Gould. Temporally consistent unbalanced optimal transport for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14618–14627, 2024. 1

[8] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1