USER-CONTROLLABLE DENSE VIDEO CAPTIONING: A LARGE-SCALE BENCHMARK AND FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense video captioning (DVC) aims to generate temporally localized captions for multiple events in untrimmed videos. Despite recent advances, existing methods produce captions at a fixed granularity because current benchmarks provide only single-style annotations, leaving variations in event density and caption depth largely unexplored. To address this gap, we present User-Controllable Captions (UC Captions), a new dataset with annotations that vary in event density (*i.e.,* how frequently events are localized) and caption depth (*i.e.,* the level of descriptive detail provided for each event). This is the first DVC dataset to explicitly encode controllable dimensions of annotation. Building on this, we propose User-Controllable DVC (UC-DVC), a framework that incorporates user-defined density and depth parameters to dynamically adjust event localization and caption generation. Direct evaluation shows that UC-DVC changes primarily along the requested control axis, produces ordered responses at unseen intermediate inputs, and preserves standard DVC performance across three backbones. These results establish UC Captions as a benchmark for evaluating controllable granularity and UC-DVC as a strong reference framework for the task. To support further research, both UC Captions and UC-DVC code will be publicly released after review.

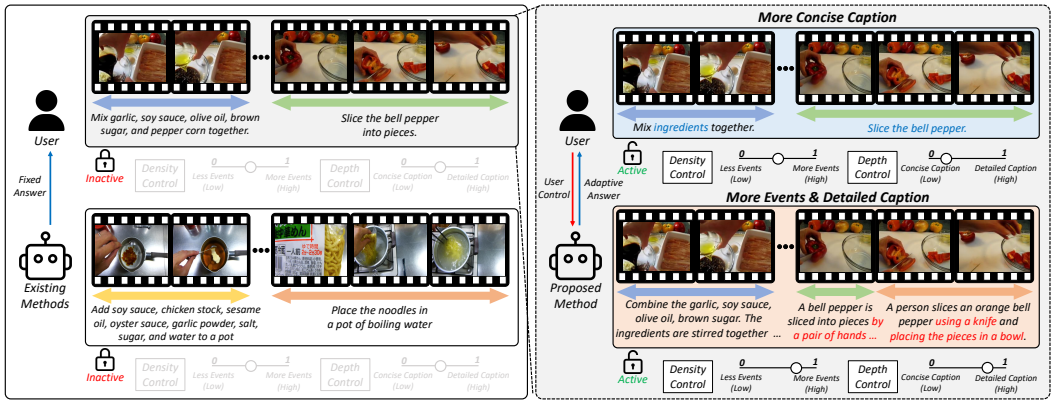
1 INTRODUCTION

Dense video captioning (DVC) has long been studied as a task that involves both temporal localization of multiple event boundaries in untrimmed videos and generating natural language descriptions for each event (Krishna et al., 2017; Zhou et al., 2018b). Unlike conventional video captioning (Chen et al., 2018; Xu et al., 2018; Zhang et al., 2020; Lin et al., 2022; Luo et al., 2020; Seo et al., 2022), which describes a single short, pre-segmented clip, DVC localizes multiple events and generates a contextually appropriate caption for each event. Recent methods have progressed from two-stage detect-then-describe pipelines to models that jointly localize and describe events (Mun et al., 2019; Iashin & Rahtu, 2020; Wang et al., 2018; Yang & Yuan, 2018; Wang et al., 2020). More recently, models pretrained at scale on narrated videos have learned rich temporal alignments and semantic representations from multimodal data (Yang et al., 2023b;a; Huang et al., 2020; Miech et al., 2019; Zellers et al., 2021; Kim et al., 2024; 2025; Jeon et al., 2025).

Despite these advancements, existing DVC methods remain limited in adapting to user-specific demands, as users may require different levels of event granularity and detail depending on their time constraints and goals. For example, surveillance officers may prefer concise summaries of suspicious activities, whereas medical trainees may require fine-grained, step-by-step descriptions of surgical procedures. Jiang et al. (2024) further showed that people with low vision prefer detailed descriptions for educational content but concise descriptions for entertainment. However, as shown in Figure 1(a), existing methods are trained on benchmarks with a single fixed annotation style and consequently generate uniform outputs, creating a persistent gap between research advances and real-world usability.

To overcome this limitation, we introduce **User-Controllable Captions (UC Captions)**, a dataset derived from existing DVC benchmarks (YouCook2 (Zhou et al., 2018a) and ViTT (Huang et al., 2020)) and enriched with annotations at multiple levels of event granularity and descriptive detail. We further propose **User-Controllable DVC (UC-DVC)**, a flexible baseline framework that adapts DVC outputs to diverse user requirements.

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.



(a) *Existing Methods* (b) ***Ours***

Figure 1: Conceptual comparison of (a) existing methods, which generate fixed captions, and (b) ours, which enables user-controllable captions by adjusting event density and caption depth.

Table 1: Comparison of UC Captions with existing datasets: (a) dataset statistics, (b) average event duration and word counts (avg. duration / avg. words), (c) segment–caption CLIP alignment

Dataset	# Captions	Annotations
ActivityNet	100K	1 Dens. & 1 Dep.
YouCook2	15K	1 Dens. & 1 Dep.
ViTT	56K	1 Dens. & 1 Dep.
UC Captions (YouCook2)	**121K**	**3 Dens. & 3 Dep.**
UC Captions (ViTT)	**337K**	**3 Dens. & 3 Dep.**

(a) Statistical comparison between datasets (Dens.:Density, Dep.:Depth).

Depth Level	Density Level (YouCook2)			Density Level (ViTT)		
	Low	Mid	High	Low	Mid	High
Low	26.03/3.12	19.58/3.09	14.76/3.09	73.55/2.98	57.31/2.94	39.82/2.84
Mid	26.03/8.40	19.58/8.77	14.76/8.38	73.55/8.11	57.31/8.08	39.82/7.97
High	26.03/18.56	19.58/18.51	14.76/18.63	73.55/17.61	57.31/17.74	39.82/17.85

(b) Average event duration (left) and word count (right) in UC Captions (underlined: annotations from the original datasets).

Condition	YouCook2	ViTT
Random (other video)	.186	.165
UC Captions	**.242**	**.226**
Human GT (anchor)	.245	.214

(c) CLIP alignment of the annotations.

First, we construct UC Captions around two key aspects of controllability, (*i*) **event density**, which determines how many events within a video are segmented and described, and (*ii*) **caption depth**, which specifies the level of descriptive detail for each event. We annotate both aspects at three levels—low, mid, and high—mapped to control values of 0, 0.5, and 1.0, respectively, thereby forming a structured and extensible representation of granularity. Training with these values enables the framework to learn a continuous control space that supports interpolation between 0 and 1 for finer control over event density and caption depth (Figure 1(b)).

Second, we introduce UC-DVC, which adaptively generates captions at user-specified levels of event density and caption depth. While UC Captions provides the necessary annotations, existing DVC architectures cannot naturally accommodate such controllable inputs. UC-DVC addresses this gap with dedicated modules that learn density and depth representations from user-defined values between 0 and 1, yielding captions that follow the requested granularity and offering a flexible baseline for user-controllable DVC aligned with real-world needs.

The major contributions of our paper are as follows.

- **UC Captions.** The first DVC benchmark that annotates each retained source video on a 3×3 grid over event density and caption depth, yielding 121K and 337K event–caption annotations on YouCook2 and ViTT where prior benchmarks provide a single fixed style.
- **UC-DVC.** A backbone-compatible framework that takes the density and depth requests as continuous inputs through dedicated modules and learns their effect from paired target-reference annotations, integrated into three DVC backbones.
- **Direct Controllability Evaluation.** UC-DVC follows the requested density and depth, stays ordered at unseen intermediate inputs, and keeps standard DVC performance on three backbones.

2 UC Captions Dataset

Following recent DVC approaches (Kim et al., 2025; Jeon et al., 2025), we use YouCook2 (Zhou et al., 2018a) and ViTT (Huang et al., 2020) as base datasets for constructing UC Captions. YouCook2 consists of 2,000 cooking-procedure videos, while ViTT contains 8,000 open-domain (general-domain) videos randomly sampled from YouTube-8M (Abu-El-Haija et al., 2016).

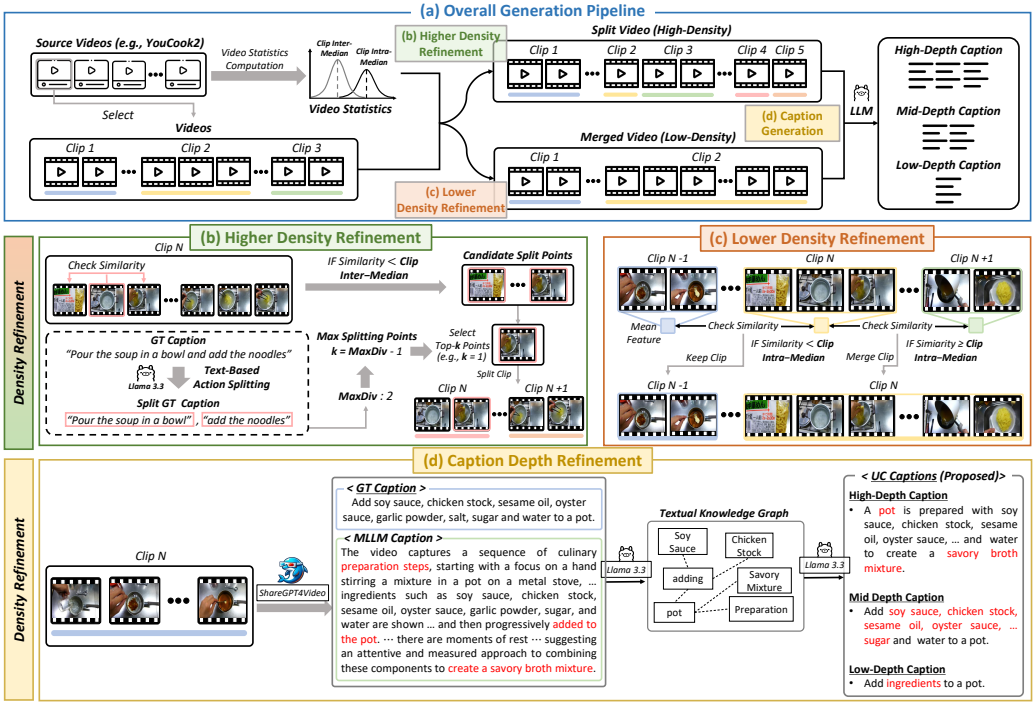


Figure 2: Generation pipeline of UC Captions. **(a) Overall Generation Pipeline** Visual similarity statistics establish data-driven thresholds for event merging (*Clip **Intra**-Median*) and splitting (*Clip **Inter**-Median*). **(b, c) Density Refinement:** Event density is controlled by merging visually similar clips or splitting incoherent ones. Splitting is further constrained by the number of actions in the ground-truth text (*MaxDiv*) to ensure semantic relevance. **(d) Caption Depth Refinement:** Caption depth is controlled by generating multi-level descriptions.

2.1 Granularity-Controlled Dataset Construction

Existing DVC benchmarks (*i.e.*, YouCook2, ViTT) are restricted to fixed levels of granularity for a given video (Table 1(a)). As a result, they cannot adequately support captioning that meets diverse user requirements. A natural alternative is to request the granularity from an MLLM directly. However, MLLM-based DVC still trails task-specific DVC even at the original annotation style, before any granularity is requested. The strongest, TimeExpert, reaches 39.0 CIDEr and 7.2 SODA$_c$ on YouCook2 against 75.80 and 10.28 for Sali4Vid (Table 2), so task-specific DVC models remain the practical choice, yet they learn from these single-style benchmarks. Natural language is also too ambiguous to separate levels, since a request such as "briefly summarize" does not state how many events to localize or how detailed each caption should be. We therefore make granularity a numeric control, extend these benchmarks into the UC Captions dataset with annotations for each control level, and train UC-DVC on it (Section 3). Specifically, we define two controllable aspects—**event density** and **caption depth**—and divide each into **three levels**, yielding a comprehensive 3×3 granularity spectrum.

We establish the varying granularity levels by analyzing the existing benchmarks. Captions in YouCook2 consist of complete sentences that describe groups of related actions (*e.g.,* "combine soy sauce, water and sugar in a bowl"). Based on this, we define the **mid-density** level as a group of actions composing a single step, and the **mid-depth** level as a complete sentence of 7-8 words. Conversely, ViTT provides concise tags for atomic actions (*e.g.,* "adding water"). This leads to our definition of the **high-density** level as a single atomic action, and the **low-depth** level as a concise tag of 2-3 words. These two benchmarks serve as data-based anchors for our spectrum. Anchoring on human

Table 2: MLLM-based DVC and Sali4Vid, a task-specific DVC model trained on the original annotation, on YouCook2 *val*.

Method	$\mathcal{C}$	$\mathcal{S}$	F1
TimeChat (Ren et al., 2024)	11.0	3.4	19.5
VTG-LLM (Guo et al., 2025a)	13.4	3.6	20.6
TRACE (Guo et al., 2025b)	35.5	6.7	31.8
TimeExpert (Yang et al., 2025)	39.0	7.2	33.5
Sali4Vid (Jeon et al., 2025)	**75.80**	**10.28**	**33.61**

Table 3: Content of UC Captions along both axes. (a) Segment–caption CLIP alignment per cell. Underlined cells are the original annotation. (b) Gemini-2.5-Flash (Comanici et al., 2025) scores (0–1) and ranks (1–3, 1 is richest) the three depths of 2,000 events per dataset.

	YouCook2 (depth)			**ViTT** (depth)		
Density	Low	Mid	High	Low	Mid	High
Low	.224	.251	.261	.206	.228	.241
Mid	.222	<u>.245</u>	.260	.205	.228	.241
High	.218	.244	.256	<u>.214</u>	.228	.239

(a) Density × depth. Segment-caption CLIP alignment.

	YouCook2		**ViTT**	
Depth	Score↑	Rank↓	Score↑	Rank↓
Low	0.37	2.98	0.26	2.81
Mid	0.74	1.93	0.56	2.18
High	0.95	1.10	0.87	1.02

(b) Depth. LLM-judged detail.

annotation keeps every cell tied to events that annotators already verified, so the pipeline refines the granularity of real events instead of inventing new ones.

The remaining levels are defined by extrapolating from these anchor points. We define the **low-density** level as a major phase that covers multiple mid-density steps, representing a broader segment of activity. The **high-depth** level is defined as a detailed narrative of 18-20 words that elaborates on a mid-depth sentence, providing richer context and explanation. This completes the formal definition of our 3×3 granularity spectrum. A detailed analysis of this spectrum is provided in Table 1.

With the full spectrum defined in this data-driven manner, we construct the complete dataset through a scalable generative approach. Leveraging advanced large language models (LLMs) and multi-modal large language models (MLLMs) (Chen et al., 2024a;b), our process systematically expands from the human-annotated anchors to populate the remaining levels. For instance, our MLLM–LLM pipeline generates **high-depth** captions by elaborating on mid-depth sentences or produces **low-density** captions by summarizing a sequence of mid-density events. This methodology enables the creation of a large-scale, diverse, and controllable dataset that is firmly grounded in real-world data characteristics.

2.2 Generation Pipeline

As shown in Figure 2, to realize the flexible granularity defined by our framework, we design a two-phase generation pipeline. This pipeline is engineered to provide precise, independent control over both event density and caption depth. Its foundation lies in a set of data-driven thresholds derived from a statistical analysis of visual similarity (*i.e.,* cosine-similarity) within our Source datasets (*i.e.,* YouCook2 or ViTT). This analysis yields two thresholds: the *Clip Intra-Median*, which captures the median similarity within a single coherent event and guides event merging, and the *Clip Inter-Median*, representing the median similarity between distinct events, which guides event splitting. This approach ensures that all granularity adjustments are contextually grounded.

Higher/Lower Density Refinement. The first phase uses these pre-established thresholds to strategically adjust temporal boundaries. (*i*) To generate higher density annotations, it splits clips at points of low internal coherence where similarity drops below *Clip Inter-Median*. To maintain semantic integrity during this splitting process, we introduce *MaxDiv*, a constraint where an LLM parses the GT caption to identify its distinct actions (Figure 2(b)), thus aligning the number of possible splits with the number of semantic actions. Conversely, (*ii*) to generate lower density annotations, the pipeline merges adjacent clips whose visual similarity exceeds the *Clip Intra-Median*, effectively grouping them into a single, broader event (Figure 2(c)). Details about *MaxDiv* and the statistical validation for *Clip Inter-/Intra- Median* can be found in Appendix B.1.

Caption Depth Refinement. The second phase generates descriptions of varying depth (Figure 2(d)). To ensure semantic consistency and high fidelity, we introduce a **Textual Knowledge Graph (TKG)**, aligning with recent work on LLM-based knowledge graph construction (Pan et al., 2024). This strategy is motivated by findings that grounding LLM generation in structured knowledge enhances reliability and mitigates hallucinations (Zhang & Soh, 2024; Zhu et al., 2024). First, we use an MLLM to generate a rich description and then construct the knowledge graph using an LLM. Next, we query the LLM with this knowledge graph to generate captions for each depth level. By selectively querying this structured knowledge, the LLM crafts descriptions ranging from concise tags (low depth) to context-rich narratives (high depth) with high precision. This is critical for achieving

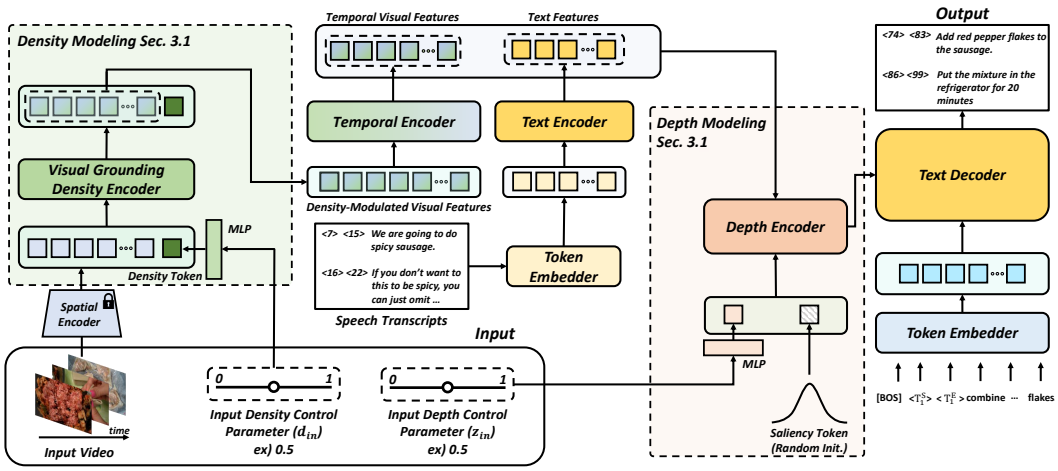


Figure 3: Network configuration of the proposed UC-DVC framework.

consistent and fine-grained control over caption content. A detailed description of the procedure and prompts is provided in Appendix B.2.

2.3 Dataset Analysis and Quality Validation

Comparison with Existing Benchmarks. Applying our pipeline (Section 2.2) to the YouCook2 *train*/*val* and ViTT *train*/*test* splits yields UC Captions, a large-scale resource with a 3-density by 3-depth annotation structure, as summarized in Table 1(a).

Quantitative Validation. As shown in Table 1(b), increasing the density control from low to high systematically decreases the average event duration (*e.g.,* from 26.03s to 14.76s on YouCook2). Conversely, increasing the depth control from low to high consistently raises the average word count (*e.g.,* from 3.12 to 18.56 words). All statistics in Table 1(b) and the similarity thresholds that drive the pipeline are computed on the training split only (Appendix B.1).

Label-Quality Calibration. We further measure the released labels directly. Table 1(c) reports segment–caption CLIP alignment over the full grid against two references. *Random* is a caption from a different video, and *Human GT* is the original anchor annotation. UC Captions stays well above the random floor on either dataset (.242 and .226 against .186 and .165) and within .012 of the human anchor, so the captions describe what their videos show.

Semantic Depth and Density. Table 3(a) breaks this alignment down by cell. It rises with the requested depth at every density level and changes by at most 0.009 across density levels at every depth, so re-segmenting keeps each caption matched to its segment. An LLM judge also scores the Low, Mid, and High captions of the same event for concrete details such as actions, objects, and tools (Table 3(b), prompts in Figures 21–22). The score rises from 0.37 to 0.95 on YouCook2 and from 0.26 to 0.87 on ViTT, so deeper levels add concrete detail to the same event.

Dataset Quality Control. To verify the naturalness of the generated captions, we conducted a human evaluation, whose protocol and results are provided in Appendix C. Experts reviewed every density–depth combination while watching the clips, and their ratings confirm both the quality of the annotations and the perceptual effectiveness of our control axes. In a blinded variant that hides the anchor, experts still identify the level of both axes far above chance (Table 17), and an entailment check shows that the generated captions rarely contradict the human annotation (Appendix B.3).

3 Proposed Baseline Framework: UC-DVC

In dense video captioning (DVC), a model temporally localizes multiple events in an untrimmed video and generates a caption for each event (Krishna et al., 2017; Wang et al., 2021). Given frame-level visual features $V = \{v_1, \ldots, v_l\} \in \mathbb{R}^{l \times h}$ and transcript features $S_t = \{t_1, \ldots, t_s\} \in \mathbb{R}^{s \times h}$, a DVC backbone predicts a timestamp–caption sequence $Z = \{(\langle T_s^n \rangle, \langle T_e^n \rangle, c_n)\}_{n=1}^{N}$.

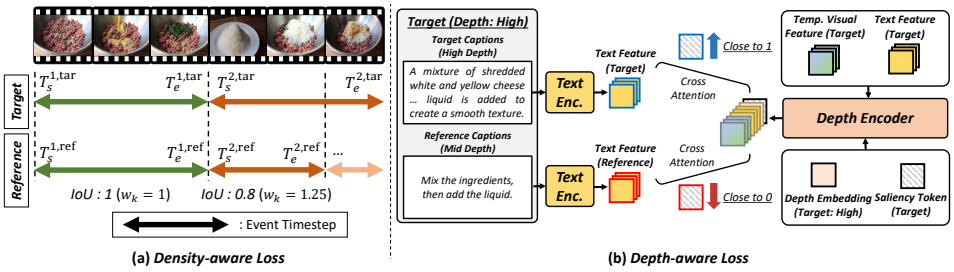


Figure 4: Explanation of proposed (a) density-aware loss and (b) depth-aware loss.

3.1 Architecture and Control Modeling

With UC Captions in place, the remaining gap is the model, which has no way to receive a user request. To overcome this limitation, UC-DVC introduces two continuous control inputs, density d_{in} and depth z_{in}, both defined in $[0, 1]$, so that the same video can be decoded at different event and caption granularities. The low, mid, and high levels of UC Captions correspond to 0, 0.5, and 1, respectively, while intermediate values can also be supplied at inference time.

As illustrated in Figure 3, the two controls are introduced at different stages according to their roles. Density determines how a video is temporally partitioned and is therefore incorporated before temporal encoding, allowing it to influence event boundaries and the number of localized events. In contrast, depth determines how much information is expressed for each event and is incorporated after temporal visual encoding, closer to the text decoder.

The two control values are projected into the shared feature space as:

$$e_d = W_d d_{\text{in}} + b_d, \qquad e_z = W_z z_{\text{in}} + b_z, \tag{1}$$

where $e_d, e_z \in \mathbb{R}^h$ are learnable control embeddings. The density embedding is combined with the frame-level visual features and processed by the density encoder:

$$V_d = E_d([V \| e_d]). \tag{2}$$

The resulting features are passed through the temporal encoder to obtain $\bar{V}_d$. The depth encoder then combines $\bar{V}_d$ with the transcript features, depth embedding, and a learnable saliency token $s \in \mathbb{R}^h$:

$$V_z = E_z([\bar{V}_d \| S_t \| s \| e_z]). \tag{3}$$

The text decoder generates the timestamp–caption sequence from V_z. This design stays compatible with existing DVC backbones while separately controlling event granularity and caption detail.

3.2 Granularity-Aware Learning Objectives

UC Captions provides multiple valid timestamp–caption annotations for the same video under different density-depth settings. To explicitly learn how predictions should change across these settings, we sample a target annotation Y_{tar} corresponding to the requested control pair $(d_{\text{tar}}, z_{\text{tar}})$ and a reference annotation Y_{ref} from a different granularity setting of the same video. Because the target and reference describe the same underlying visual content, their differences provide direct supervision for granularity variation (Figure 4).

Density-Aware Sequence Loss. We denote the tokenized target timestamp–caption sequence by $\mathbf{y} = (y_1, \ldots, y_L)$ and optimize the following density-aware sequence-generation objective:

$$\mathcal{L}_\theta(V_z, \mathbf{y}) = -\frac{1}{\sum_{k=1}^{L} w_k} \sum_{k=1}^{L} w_k \log p_\theta(y_k \mid V_z, y_{<k}). \tag{4}$$

For each target event i, IoU_i is the mean of its maximum temporal IoUs with annotated events at each other density level of the same video, which is always positive because all density levels are derived from the same anchor events (Section 2.2). For a timestamp token y_k associated with event i, we assign $w_k = 1/\text{IoU}_i$, while non-timestamp tokens retain unit weight. Timestamp tokens corresponding to larger boundary differences therefore receive greater emphasis.

Table 4: Axis-isolated control sweep with the Sali4Vid backbone. Cells report events/video and words/event, and **bold** marks the requested quantity. Original and UC Captions rows are annotation statistics. Inputs .25 and .75 are unseen during training.

Model	Sweep	YouCook2 (*val*)					ViTT (*test*)				
		0	**.25**	**.5**	**.75**	**1**	**0**	**.25**	**.5**	**.75**	**1**
Original	*fixed*	–	–	7.68/8.71	–	–	–	–	–	–	7.13/3.06
UC Captions	d (z=.5)	6.26/8.38	–	7.68/8.71	–	9.94/8.39	2.75/8.28	–	4.27/8.04	–	7.13/7.96
	z (d=.5)	7.68/3.13	–	7.68/8.71	–	7.68/18.49	4.27/2.99	–	4.27/8.04	–	4.27/17.77
UC-DVC	d (z=.5)	**5.24**/8.50	**6.45**/8.49	**7.66**/8.48	**8.81**/8.47	**9.94**/8.45	**3.70**/8.30	**3.95**/8.29	**4.78**/8.20	**6.00**/8.06	**7.33**/7.93
	z (d=.5)	7.66/**3.12**	7.66/**5.50**	7.66/**8.48**	7.66/**12.51**	7.66/**16.56**	4.51/**2.97**	4.77/**8.01**	4.78/**8.20**	4.53/**15.64**	4.46/**16.56**

Table 5: Nine-cell comparison with Sali4Vid. One UC-DVC, trained on all cells, and nine Sali4Vid models, each trained on one cell, are scored on each cell's annotation as Sali4Vid / **UC-DVC** (gain).

Density	Metric	UC Captions (YouCook2)			UC Captions (ViTT)		
		Low Depth	**Mid Depth**	**High Depth**	**Low Depth**	**Mid Depth**	**High Depth**
Low	$\mathcal{M}$	9.02 / **11.21** (+2.19)	9.64 / **11.34** (+1.70)	6.60 / **9.26** (+2.66)	2.38 / **4.89** (+2.51)	2.47 / **9.14** (+6.67)	1.41 / **4.92** (+3.51)
	F1	30.98 / **32.68** (+1.70)	31.07 / **32.89** (+1.82)	31.04 / **32.34** (+1.30)	33.56 / **36.12** (+2.56)	31.75 / **38.43** (+6.68)	29.56 / **33.91** (+4.35)
Mid	$\mathcal{M}$	9.57 / **12.47** (+2.90)	13.57 / **14.02** (+0.45)	6.71 / **10.38** (+3.67)	3.90 / **7.04** (+3.14)	3.68 / **12.75** (+9.07)	1.91 / **5.93** (+4.02)
	F1	31.75 / **34.97** (+3.22)	33.79 / **36.91** (+3.12)	32.44 / **35.04** (+2.60)	43.33 / **43.55** (+0.22)	42.99 / **43.68** (+0.69)	41.16 / **44.72** (+3.56)
High	$\mathcal{M}$	8.21 / **11.99** (+3.78)	9.29 / **12.26** (+2.97)	6.06 / **9.86** (+3.80)	9.82 / **12.39** (+2.57)	3.91 / **12.11** (+8.20)	2.04 / **6.85** (+4.81)
	F1	28.14 / **32.54** (+4.40)	29.30 / **33.36** (+4.06)	28.13 / **30.90** (+2.77)	45.89 / **47.62** (+1.73)	45.74 / **46.55** (+0.81)	45.75 / **46.54** (+0.79)

Depth-Aware Comparative Loss. To supervise caption depth, we let V_z cross-attend to the encoded target or reference caption and map the output at the saliency-token position, V_s, to a matching score:

$$S = \sigma(W_s V_s + b_s). \tag{5}$$

We assign $r = 1$ to the target caption and $r = 0$ to the reference caption, taken at a different depth of the same video, and optimize

$$\mathcal{L}_s = -r \log S - (1 - r) \log(1 - S). \tag{6}$$

This same-video comparative objective encourages the conditioned representation to distinguish the caption detail of the requested depth level rather than relying only on the video content.

Training Objective. The final training objective is

$$\mathcal{L} = \mathcal{L}_\theta + \lambda \mathcal{L}_s, \tag{7}$$

where the density-aware sequence loss emphasizes changes in temporal localization and the depth-aware comparative loss supervises caption detail. We set $\lambda = 1$ in all experiments, and performance is insensitive to moderate variations in its value.

4 Experiments

4.1 Experimental Settings

Evaluation Metrics. We evaluate performance using CIDEr ($\mathcal{C}$) (Vedantam et al., 2015) and METEOR ($\mathcal{M}$)(Banerjee & Lavie, 2005) to evaluate the captioning quality, following (Yang et al., 2023b). For event localization, we follow the standard protocol by computing the F1 score, the harmonic mean of average precision and recall, across IoU thresholds of $\{0.3, 0.5, 0.7, 0.9\}$. Additionally, since these metrics do not fully reflect the overall caption coherence of dense video captioning, we also employ SODA_c ($\mathcal{S}$) (Fujita et al., 2020) to evaluate overall storytelling performance.

Implementation Details. Following Vid2Seq (Yang et al., 2023b), we sample video frames at 1 fps on YouCook2 (Zhou et al., 2018a) and ViTT (Huang et al., 2020). The text encoder and decoder are initialized from T5-Base (Raffel et al., 2020), and CLIP ViT-L/14 (Radford et al., 2021) is used as the visual encoder. Additional optimization and training details are provided in Appendix D.

Table 6: Standard DVC performance at the original operating points, $(d, z) = (0.5, 0.5)$ for YouCook2 and $(1, 0)$ for ViTT. Published results are included for context, and † denotes our reproductions. *Full UC–Uncond.* uses the same multi-granularity annotations without control inputs, whereas UC-DVC receives the requested (d, z).

Method	Fine-tune labels	Interface	YouCook2 (*val*)				ViTT (*test*)			
			$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1
Vid2Seq (Yang et al., 2023b)	Source	Fixed	12.30	67.20	10.30	32.60	9.50	50.00	15.00	46.20
HiCM^2 (Kim et al., 2025)	Source	Fixed	12.80	71.84	10.73	32.51	9.66	51.29	15.07	45.08
Sali4Vid (Jeon et al., 2025)	Source	Fixed	13.54	75.80	10.28	33.61	10.05	53.87	15.08	46.58
Vid2Seq†	Source	Fixed	12.25	67.51	10.30	32.17	9.41	48.85	14.19	44.93
Vid2Seq†	Full UC	Uncond.	8.30(-3.95)	41.59(-25.92)	8.42(-1.88)	30.96(-1.21)	4.90(-4.51)	10.40(-38.45)	2.61(-11.58)	43.15(-1.78)
Vid2Seq + UC-DVC	Full UC	(d, z)	**12.81**(+0.56)	**73.14**(+5.63)	**10.81**(+0.51)	**33.11**(+0.94)	**11.14**(+1.73)	**51.62**(+2.77)	**15.49**(+1.30)	**46.56**(+1.63)
HiCM^2†	Source	Fixed	12.52	68.92	10.71	32.54	9.37	49.08	15.01	44.92
HiCM^2†	Full UC	Uncond.	8.38(-4.14)	43.72(-25.20)	8.40(-2.31)	31.00(-1.54)	5.29(-4.08)	15.27(-33.81)	2.34(-12.67)	42.17(-2.75)
HiCM^2 + UC-DVC	Full UC	(d, z)	**12.92**(+0.40)	**75.97**(+7.05)	**11.91**(+1.20)	**37.90**(+5.36)	**11.28**(+1.91)	**53.69**(+4.61)	**15.58**(+0.57)	**46.49**(+1.57)
Sali4Vid†	Source	Fixed	13.57	76.13	10.41	33.79	9.82	52.59	15.20	45.89
Sali4Vid†	Full UC	Uncond.	8.95(-4.62)	44.79(-31.34)	7.21(-3.20)	29.65(-4.14)	3.49(-6.33)	12.53(-40.06)	5.93(-9.27)	42.71(-3.18)
Sali4Vid + UC-DVC	Full UC	(d, z)	**14.02**(+0.45)	**82.35**(+6.22)	**11.23**(+0.82)	**36.91**(+3.12)	**12.39**(+2.57)	**55.34**(+2.75)	**15.97**(+0.77)	**47.62**(+1.73)

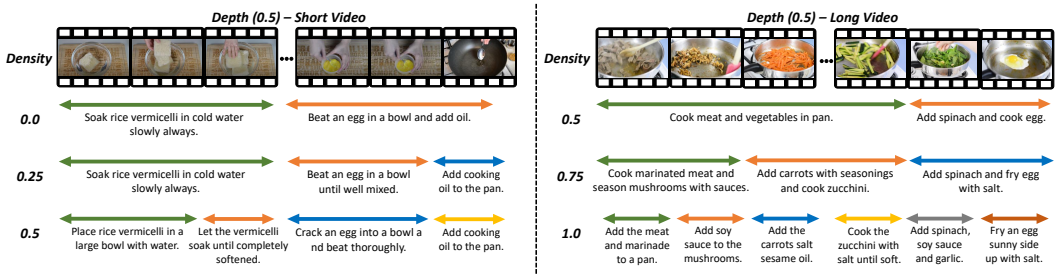


Figure 5: Visualization results for intermediate density levels on the YouCook2 *val* set. The figure shows outputs for density values of 0.25 and 0.75, neither of which is seen during training.

4.2 Direct Evaluation of Controllability

Axis-Isolated Adherence. A controllable model should change the requested output property while keeping the non-target axis stable. We therefore vary density and depth independently at inference time while fixing the other control to 0.5. As shown in Table 4, increasing density on YouCook2 raises the number of predicted events from 5.24 to 9.94, while caption length remains nearly unchanged from 8.50 to 8.45 words per event. Conversely, increasing depth expands captions from 3.12 to 16.56 words per event, while the event count remains fixed at 7.66. ViTT, an open-domain dataset drawn from YouTube-8M, exhibits the same axis separation. Density changes events per video from 3.70 to 7.33 with only a 0.37-word variation, whereas depth changes words per event from 2.97 to 16.56 with only a 0.32-event variation. The responses also remain ordered at the unseen inputs 0.25 and 0.75 (Table 4 and Figure 5), and on ViTT the mean per-video Spearman correlation over the five inputs is .897 for density and .948 for depth, so the ordering holds within individual videos.

Content at the Requested Granularity. Beyond counts and lengths, we score the outputs at the annotated levels against the annotation of the requested cell and compare them with Sali4Vid trained only on that cell (Table 5), the stronger baseline since pooling all cells without control degrades Sali4Vid (Table 6). UC-DVC scores higher in all nine cells of both datasets in METEOR and F1, and deeper captions add actions, tools, and setting (Figure 8 in Appendix F).

4.3 Preserving Standard DVC Performance

Controllability should not come at the cost of standard DVC performance. We therefore evaluate UC-DVC at the original operating point of each benchmark. As shown in Table 6, UC-DVC improves every reported metric over the corresponding source-trained reproduction across all three backbones and both datasets, showing that flexible density and depth control does not sacrifice event localization or captioning quality. In contrast, training on the same Full UC annotations without control inputs substantially degrades performance because the multiple granularity-specific targets associated with each video become indistinguishable.

Table 7: Effect of UC Captions generation components on YouCook2 *val* set.

Data Type	MaxDiv	TKG	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1
Original	-	-	13.57	76.13	10.41	33.79
UC Captions	✗	✗	13.64	77.18	10.74	35.12
	✓	✗	13.71	79.42	10.53	35.04
	✗	✓	13.86	80.05	10.73	36.48
	✓	✓	**14.02**	**82.35**	**11.23**	**36.91**

Table 8: Effect of proposed UC-DVC components on YouCook2 *val* set.

Density	Depth	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1
✗	✗	8.95	44.79	7.21	29.65
✓	✗	11.74	60.87	9.98	35.92
✗	✓	13.28	76.53	10.21	34.82
✓	✓	**14.02**	**82.35**	**11.23**	**36.91**

Table 9: Generation cost of UC Captions on the available videos (of 2,000 and 8,000).

Dataset	# Videos	GPU Hours	Wall-clock (h)
YouCook2	1,686	≈ 69	≈ 34.5
ViTT	7,461	≈ 306	≈ 153.0
Total	**9,147**	$\approx$ **375**	$\approx$ **187.5**

Table 10: Comparisons of training time, inference time, and number of parameters.

Method	Train (s) *(per iter)*	Infer (s) *(per video)*	# Params
Vid2Seq (Yang et al., 2023b)	0.208	0.681	289.2M
Vid2Seq+Ours	0.229	0.705	324.7M
HiCM^2 (Kim et al., 2025)	0.941	0.904	290.3M
HiCM^2+Ours	0.986	0.933	325.8M
Sali4Vid (Jeon et al., 2025)	0.163	0.924	288.4M
Sali4Vid+Ours	0.180	0.955	323.9M

4.4 Ablation Study

We conduct ablation studies on UC Captions (YouCook2) using Sali4Vid baseline to investigate (1) the effect of UC Captions generation components and (2) the effect of UC-DVC modeling components. A detailed analysis of the ablations and generalization is provided in Appendix E.

Effect of UC Captions Generation Components. We first investigate the effect of components in UC Captions generation. Table 7 reports downstream quality. *MaxDiv* and TKG each improve over the ablated pipeline and the two together perform best, so the components contribute complementary gains rather than substituting for each other.

Effect of UC-DVC Modeling Components. Table 8 shows the effect of the two UC-DVC components: density modeling and depth modeling. Both components improve performance, reflecting their distinct roles in event localization and captioning. Keeping both control modules but turning off both objectives, with unit timestamp weights in place of density weighting, gives up about half of the METEOR gain over the model without controls (Table 19 in Appendix E).

4.5 Discussion

Computational Costs. Generating UC Captions for all $9{,}147$ available videos took about 375 GPU-hours, or about 187.5 hours of wall-clock time on two RTX A6000 GPUs (Table 9). Each video needs three density passes to produce all nine density–depth combinations. On the model side, the two control inputs add 35.5M parameters to each backbone, increase training time by 5–10%, and increase inference time by about 3% (Table 10). Given the advantage of dynamically handling user-controllable density and depth, we believe this additional cost is marginal for practical scenarios.

Limitations. Anchoring on human annotation keeps the grid free of invented events, and also fixes the scope of the benchmark. The pipeline refines annotated intervals rather than creating them, so a higher density request subdivides annotated events and cannot reach unannotated stretches. This is the right trade for a benchmark whose value rests on trustworthy labels, and a wider scope needs wider human annotation first. Beyond the two anchor cells, the grid comes from our pipeline, validated by entailment, grounding, and an anchor-blinded study rather than human re-annotation.

5 Conclusion

We introduced user-controllable dense video captioning, in which event density and caption depth are explicit inputs rather than fixed properties of a benchmark. UC Captions provides a 3×3 annotation grid over these axes on YouCook2 and ViTT, while UC-DVC uses axis-specific conditioning and paired target–reference objectives to learn the requested granularity. Direct evaluation shows ordered, largely axis-specific responses, including at unseen intermediate inputs, while standard DVC performance is preserved across three backbones. Together, the benchmark and framework establish a concrete foundation for evaluating and improving controllable granularity in DVC.

AI USE STATEMENT

Large language and multimodal models are part of the UC Captions construction pipeline (Section 2.2 and Appendix B), and Gemini-2.5-Flash serves as an automatic judge of caption depth (Section 2.3 and Appendix F.10). We used ChatGPT and Claude solely to assist with manuscript drafting and editing. The authors reviewed and revised all AI-assisted text and take responsibility for the final content.

ETHICS STATEMENT

UC-DVC gives users explicit control over event density and caption depth, which can serve accessibility needs, from coarse summaries for quick scanning to detailed descriptions. UC Captions also offers a controlled testbed for studying over- and under-segmentation, hallucination, and bias across granularities.

Risks and mitigation. Like other video understanding systems, UC-DVC could be misused to produce misleading narratives or to support surveillance-style analytics. We therefore release only annotations and code, never the original videos, and license UC Captions for research use only, prohibiting commercial use, re-identification, and any attempt to link annotations to individuals.

Data provenance and licensing. UC Captions builds on the public YouCook2 and ViTT datasets under their terms. The release contains annotations that reference public video identifiers, and the captions generated with ShareGPT4Video and Llama 3.3-70B are marked as derivative research artifacts. Generated captions can inherit errors and biases of these models, and we audit them with entailment, grounding, and human checks (Section 2.3).

Privacy and legal compliance. We do not store personally identifiable information, and we comply with applicable copyright and data-protection laws.

Human subjects. The human evaluation (Appendix C) was conducted with informed consent, clear task instructions, fair compensation, and minimal data collection, following institutional ethics review processes where required. No sponsor influenced the study design, analysis, or publication decisions.

REPRODUCIBILITY STATEMENT

Section 2.2 and Appendix B describe the annotation pipeline, including the data-driven similarity thresholds for density refinement, the Textual Knowledge Graph for depth refinement, and the generation prompt. Appendix D gives the architecture, training hyperparameters, and computing environment of UC-DVC, and Appendix C gives the human evaluation protocol. The data generation code accompanies this submission as supplementary material, and the full code and UC Captions will be released upon acceptance (Appendix G).

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, 2005.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, 2024a.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, 2024b.

Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In *ECCV*, 2018.

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.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *EMNLP*, 2023.

Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. Soda: Story oriented dense video captioning evaluation framework. In *ECCV*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *AAAI*, 2025a.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. In *ICLR*, 2025b.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *CVPR*, 2024.

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

Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *BMVC*. BMVA Press, 2020.

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. In *EMNLP*, 2025.

Lucy Jiang, Crescentia Jung, Mahika Phutane, Abigale Stangl, and Shiri Azenkot. "it's kind of context dependent": Understanding blind and low vision people's video accessibility preferences across viewing scenarios. In *CHI*, 2024.

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 *CVPR*, 2024.

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Hicm^2: Hierarchical compact memory modeling for dense video captioning. In *AAAI*, 2025.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, 2022.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, 2019.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: advancing video large language model with fine-grained temporal reasoning. In *ICML*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, 2024.

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *CVPR*, 2022.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018.

Teng Wang, Huicheng Zheng, Mingjing Yu, Qian Tian, and Haifeng Hu. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1890–1900, 2020.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *ICCV*, 2021.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024a.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawk-eye: Training video-text llms for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024b.

Kangyi Wu, Pengna Li, Jingwen Fu, Yizhe Li, Yang Wu, Yuhan Liu, Jinjun Wang, and Sanping Zhou. Event-equalized dense video captioning. In *CVPR*, 2025.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

Ning Xu, An-An Liu, Yongkang Wong, Yongdong Zhang, Weizhi Nie, Yuting Su, and Mohan Kankanhalli. Dual-stream recurrent neural network for video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2482–2493, 2018.

Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vidchapters-7m: Video chapters at scale. In *NeurIPS*, 2023a.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023b.

Dali Yang and Chun Yuan. Hierarchical context encoding for events captioning in videos. In *ICIP*. IEEE, 2018.

Zuhao Yang, Yingchen Yu, Yunqing Zhao, Shijian Lu, and Song Bai. Timeexpert: An expert-guided video llm for video temporal grounding. In *ICCV*, 2025.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, 2021.

Bowen Zhang and Harold Soh. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. In *EMNLP*, 2024.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.

Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, 2020.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018a.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018b.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *CVPR*, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58, 2024.

A Related Work

A.1 Dense Video Captioning

Dense Video Captioning involves detecting events in untrimmed videos and generating captions for them (Krishna et al., 2017; Zhou et al., 2018b). Traditional methods first localize events and then generate captions separately (Mun et al., 2019; Iashin & Rahtu, 2020; Wang et al., 2018; Yang & Yuan, 2018; Wang et al., 2020). This pipeline does not account for the interaction between the two tasks. To address this, recent studies explore end-to-end models that learn both tasks together. PDVC (Wang et al., 2021) formulates captioning as a set prediction task using DETR (Zhu et al., 2020) to predict event intervals and captions simultaneously. Vid2Seq (Yang et al., 2023b) treats it as a sequence-to-sequence problem, generating timestamps and captions in a single output while incorporating speech-to-text.

Other works further refine the process. One approach introduces visual memory and streaming decoding (Zhou et al., 2024), while others address specific biases. For instance, E^2DVC (Wu et al., 2025) tackles the problem of temporal bias, where models overlook short-duration events, by using a non-trainable clustering module to ensure all potential events are treated equally. Another line of research enhances captioning through retrieval-augmented generation. CM^2 (Kim et al., 2024) first introduced this idea, and $HiCM^2$ (Kim et al., 2025) further develops it by constructing a hierarchical compact memory inspired by human cognition to retrieve more relevant knowledge. Sali4Vid (Jeon et al., 2025) injects timestamp supervision into the visual stream via sigmoid-based saliency reweighting and retrieves captions with semantic-transition-aware adaptive segmentation.

Furthermore, recent advances in multimodal large language models improve video understanding. TimeChat (Ren et al., 2024) introduces a timestamp-aware encoder and a sliding Q-Former, while VTimeLLM (Huang et al., 2024) applies boundary-aware training to refine event localization. Momentor (Qian et al., 2024) equips Video-LLMs with fine-grained temporal reasoning via a Temporal Perception Module that builds a continuous temporal token space for precise timestamp positioning. HawkEye (Wang et al., 2024b) improves temporal video grounding by adopting a coarse-grained textual segment representation and refining it with recursive grounding. VTG-LLM (Guo et al., 2025a) enhances timestamp localization by integrating timestamp knowledge into visual tokens, adding absolute-time tokens, and using slot-based token compression to support many sampled frames. TRACE (Guo et al., 2025b) models videos as sequences of events (timestamps, saliency scores, captions) and implements this via task-interleaved generation with separate encoders/heads and adaptive head switching. TimeExpert (Yang et al., 2025) further reduces task interference with a Mixture-of-Experts decoder that routes different task tokens to specialized experts. Despite these improvements, existing methods assume a fixed level of detail in captions, limiting flexibility. To address this, we introduce a synthetic dataset that reflects variable event density and caption depth and propose a novel framework that dynamically adjusts captions based on user specifications.

A.2 Synthetic Data Generation

Synthetic data generation with LLMs and MLLMs has recently received significant attention as a way to produce high-quality training data, improve model performance, and reduce the need for manual annotation (Zhang et al., 2024; Ding et al., 2023; Wang et al., 2022; Xu et al., 2023). Notably, ShareGPT4V (Chen et al., 2024a) introduces a large-scale dataset of 1.2M detailed image captions generated using GPT-4V, significantly enhancing multimodal alignment in large vision-language models. Similarly, ShareGPT4Video (Chen et al., 2024b) employs a differential sliding-window captioning strategy to generate temporally precise and context-rich video captions, improving both video understanding and generation tasks. These approaches have demonstrated substantial performance gains across multiple benchmarks, highlighting the effectiveness of synthetic data in multimodal learning. Given that DVC requires annotation of untrimmed video sequences, manual labeling is particularly expensive and time-consuming. To this end, we introduce UC Captions specifically designed for controllable density and depth variations, leveraging MLLMs and LLMs to enable more flexible and adaptable dense video captioning. Our dataset facilitates systematic evaluation across multiple granularity levels, addressing a key limitation of existing DVC benchmarks and fostering progress toward user-controllable video understanding.

Table 11: Percentiles of the intra-group (for merging) and inter-group (for splitting) similarity distributions, on the training split and on the combined splits. The two medians are the thresholds, and their separation holds in every row.

Dataset (split)	Intra-Group		Inter-Group	
	10-th Perc.	Median	Median	90-th Perc.
YouCook2 (*train*)	0.8438	0.9533	0.7619	0.8962
YouCook2 (*train+val*)	0.8447	0.9534	0.7630	0.8963
ViTT (*train*)	0.8435	0.9540	0.7922	0.9255
ViTT (*train+test*)	0.8463	0.9547	0.7977	0.9292

B Additional Details for UC Captions

This section provides detailed implementation specifics and statistical validation for the core mechanisms of our data generation pipeline, organized thematically.

B.1 Density Refinement

This subsection details the mechanisms used to control event density, including the data-driven thresholds and text-aware constraints, along with their statistical validation.

Visual Similarity Thresholds. As mentioned in the main paper (Sec. 2.2), the foundation of our pipeline lies in data-driven visual similarity thresholds. These are established by analyzing the cosine similarity of CLIP features from our anchor datasets. "Intra-Group" pairs are adjacent clips from the same ground-truth event, while "Inter-Group" pairs put a clip of that event against any clip outside it, whether in another event or in an unannotated stretch. The median similarity of each group is calculated to establish the *Clip Intra-Median* (for merging) and the *Clip Inter-Median* (for splitting).

As detailed in Table 11, the two distributions separate cleanly on both datasets. The Inter-Group median falls below the 10th percentile of the Intra-Group distribution, and the Intra-Group median rises above the 90th percentile of the Inter-Group one. Each median therefore sits outside the primary range of the other distribution, which is what makes them reliable thresholds for merging and splitting.

Split Sensitivity. The thresholds are computed on the training split. The lower row of each dataset in Table 11 recomputes them on the combined splits. The two medians move by at most .001 on YouCook2 and .006 on ViTT, no other percentile moves further, and the separation survives in every row. Table 12 repeats the exercise for the grid that Table 1(b) of the main paper reports on the training split. Every YouCook2 cell moves by at most .04 seconds and .01 words. On ViTT, the Low and Mid density durations rise from 73.55 and 57.31 seconds to 81.41 and 60.05 seconds because the ViTT evaluation videos run longer, while the High density duration moves by .03 seconds. Word counts move by at most .09 anywhere.

MaxDiv. To maintain semantic integrity during the splitting, we introduce *MaxDiv*, a text-aware constraint, as mentioned in the main paper (Sec. 2.2). While low visual similarity suggests a potential boundary for a new event, relying on visual cues alone can produce meaningless segments. The *MaxDiv* constraint prevents this by ensuring that any temporal split is grounded in the semantic content of the source caption.

This process is directly illustrated in the main paper (Fig. 2(b)). We first perform **Text-Based Action Splitting**, where a powerful LLM (Llama 3.3-70B) parses the original GT Caption into its constituent action phrases. For example, the caption "*Pour the soup in a bowl and add the noodles*" is decomposed into two distinct actions: "*Pour the soup in a bowl*" and "*add the noodles*". The number of these resulting phrases sets the value of *MaxDiv* (in this case, $\mathit{MaxDiv} = 2$). From this, we determine the maximum number of visual splits to perform, $k = \mathit{MaxDiv} - 1 = 1$. The pipeline then analyzes the video clip to identify candidate split points based on visual dissimilarity,

Table 12: Table 1(b) recomputed on the combined splits, YouCook2 *train+val* and ViTT *train+test*, as event duration (s) / words per event. Underlined cells hold the original annotation.

Depth Level	Density Level (YouCook2)			Density Level (ViTT)		
	Low	Mid	High	Low	Mid	High
Low	26.01/3.12	19.59/3.10	14.80/3.10	81.41/3.02	60.05/2.96	<u>39.79/2.91</u>
Mid	26.01/8.40	<u>19.59/8.76</u>	14.80/8.38	81.41/8.15	60.05/8.07	39.79/7.97
High	26.01/18.55	19.59/18.50	14.80/18.63	81.41/17.70	60.05/17.75	39.79/17.87

Table 13: One-way ANOVA of event duration across density levels on the training split.

Dataset	F-statistic	P-value
YouCook2	818	$p < 0.001$
ViTT	3031	$p < 0.001$

and selects up to the top-k points to perform the split. It is important to note that *MaxDiv* serves as an upper bound, not a strict target. If the text describes multiple actions that occur within a visually seamless sequence with few candidate split points, the pipeline will perform fewer than k splits, prioritizing the visual evidence. This method ensures that the final visual segments directly correspond to the semantic actions identified by the LLM.

Validation of Generated Density Levels. To validate that our density refinement process creates statistically distinct groups, we performed a one-way ANOVA test. For this test, we formulated a null hypothesis (H_0) which posits that there is no meaningful difference in the mean event durations across the Low, Mid, and High density levels. As summarized in Table 13, the extremely low p-values ($p<0.001$) lead to a strong rejection of this null hypothesis. This result confirms that our density control mechanism is effective, producing event segmentations that are demonstrably and significantly different from one another.

Lower-Density Construction for ViTT. The human annotation of ViTT serves as the High-density/Low-depth anchor. We apply the lower-density refinement once to obtain Mid density, and once more to the resulting sequence to obtain Low density. Each pass merges only adjacent annotated events and never introduces an event in an unannotated interval.

B.2 Caption Depth Refinement

This subsection details our approach for generating multi-level captions using a Textual Knowledge Graph (TKG).

Textual Knowledge Graph (TKG) for Caption Depth Refinement. To ensure semantic consistency and high fidelity across all generated caption levels, the second phase of our pipeline, Caption Depth Refinement, centers on a **TKG**. It is motivated by recent advancements showing that grounding LLM generation in structured knowledge enhances reliability and reduces issues like hallucination (Pan et al., 2024; Zhang & Soh, 2024; Zhu et al., 2024). Our process utilizes a combination of state-of-the-art models. We first employ **ShareGPT4Video** (Chen et al., 2024b) as the primary MLLM to generate an initial, rich synthetic caption, followed by **Llama 3.3-70B** (Grattafiori et al., 2024) as the primary LLM to refine this output into three distinct levels of detail. All data generation processes were conducted using one NVIDIA RTX A6000 GPU for each model.

The refinement process begins when the LLM builds the TKG using the original **GT Caption** as an anchor. The LLM first analyzes the GT Caption to extract its core actions and entities. It then uses information from the more detailed **MLLM Caption** to add supplementary details. The LLM constructs the final TKG by adding the details from the MLLM caption to the core facts from the GT, following the instruction to "**Build a Knowledge Graph (KG)**". This approach is used to control for potential MLLM hallucinations while still using its descriptive information.

For example, given the GT Caption "*fry some minced garlic roasted red chili...*", the LLM identifies 'frying' as a core action. It then adds details from the MLLM Caption, such as that the garlic is fried "*until golden*" and the sauce is "*thickening*". This results in a KG containing detailed and grounded facts, such as '("minced garlic", "frying", "golden")'.

Once this event-specific knowledge graph is constructed, the final caption depth refinement becomes a structured querying task. For instance, to generate a **High Depth** caption, the LLM is prompted to

Generation Prompt (YouCook2)

System Instruction : You are an advanced AI model specialized in Video Captioning. Your task is to generate three different levels of captions from a given Ground Truth (GT) caption and Multimodal-Large-Language Model (MLLM) caption.

Task Instructions
1. Analyze the GT caption to extract the core actions and objects.
2. Extract key information from the MLLM caption to enrich the GT caption.
3. Based on the GT Caption, build a Knowledge Graph (KG) to represent the relationships between entities and actions.
4. Generate captions at three different levels by querying the KG:
- Low Depth: Generate a 2–3 word tag representing the core action and object.
- Mid Depth:
  - Case 1: If "(Merged)" or "(Part)" is in the GT caption, generate a 7–8 word sentence describing the primary interaction.
  - Case 2: Otherwise, return the original GT caption.
- High Depth: Generate a 18–20 word sentence incorporating all relevant entities and relationships.

Few-shot Examples

Example 1 (Merged or Part in GT caption)

GT Caption: "fry some minced garlic roasted red chili and the sauce and add in the corn starch and water mixture (merged)"
MLLM Caption: "The video captures a culinary scene where garlic is fried until golden, and then red chili peppers are added to heat through. Corn starch and water are stirred into a pot of sauce, thickening it as it cooks."

KG (Knowledge Graph)
[
{"minced garlic", "frying", "golden"},
{"roasted red chili", "frying", "heat through"},
{"sauce", "corn starch", "water", "stirring", "pot", "thickening", "cooking"}
]

Output:
Low Depth : "Fry ingredients."
Mid Depth : "Fry minced garlic and red chili, then add sauce."
High Depth : "Minced garlic is fried until golden before adding roasted red chili peppers, sauce mixture is then combined with cornstarch and water."

Example 2 (No Merged or Part in GT caption):
...

Figure 6: Generation prompt for UC Captions (YouCook2).

use more details from the graph, while a **Low Depth** caption uses only the most essential actions and entities. This TKG-based method provides several critical advantages by ensuring semantic consistency since all levels are derived from the same facts, offering precise control by turning generation into a structured query, and enhancing reliability by grounding the output of the LLM to reduce the risk of hallucination. Figure 6 presents the detailed prompt, including the exact instructions and few-shot examples given to the LLM, which ensures the clarity and reproducibility of our approach.

Table 14: Contradiction rate (%) against the overlapping human annotation. *Premise* is the side taken as given.

Dataset	Premise	Low	Mid	High
YouCook2	annotation	1.8	0.4	0.2
	generated	1.0	0.4	1.2
ViTT	annotation	0.4	0.6	0.4
	generated	2.0	0.6	**3.4**

Table 15: Visual and temporal grounding over all training events (mean±std). Each depth row pools the three density levels.

Dataset	Condition	CLIPScore	tIoU
YouCook2	Random caption	$.186 \pm .031$	$.078 \pm .220$
	Low depth	$.221 \pm .032$	$.262 \pm .367$
	Mid depth	$.246 \pm .035$	$.321 \pm .383$
	High depth	$.259 \pm .034$	$.333 \pm .387$
	Human GT (anchor)	$.245 \pm .034$	$.339 \pm .384$
ViTT	Random caption	$.165 \pm .034$	$.122 \pm .261$
	Low depth	$.209 \pm .040$	$.284 \pm .368$
	Mid depth	$.228 \pm .041$	$.305 \pm .376$
	High depth	$.240 \pm .041$	$.349 \pm .388$
	Human GT (anchor)	$.214 \pm .040$	$.232 \pm .340$

B.3 Annotation Faithfulness and Grounding

Everything in this subsection is computed on the training splits of UC Captions, and no evaluation video contributes to any number reported here.

Textual Faithfulness. A generated caption should not assert what the human annotation denies. We test this without the video, by pairing every generated caption with the annotation it temporally overlaps (tIoU $\geq .1$) and classifying the pair with an entailment model (RoBERTa-large fine-tuned on MNLI, 500 events per cell) (Table 14). We report the contradiction rate, which is the outcome that matters for faithfulness, since it marks a statement the annotation rules out. Contradictions are rare on both datasets and in both directions. They peak at 1.8% on YouCook2 and at 3.4% on ViTT, the latter at high depth, where a short tag leaves an elaboration the most room to overreach. A textual test cannot see detail that the annotation never mentioned, which is the gap the grounding measurement below closes.

Visual Grounding and Temporal Fidelity. We measure grounding directly against the video over a census of every training event (Table 15). *CLIPScore* is the cosine between the CLIP text embedding of a caption and the mean-pooled 1 fps CLIP ViT-L/14 features of its span, and *tIoU* is the overlap between the annotated span and the best caption-matched window of the same width.

Two references frame the UC Captions rows. *Random caption* pairs a span with a caption from a different video, which is the score a caption earns by naming the right domain and nothing more. *Low*, *Mid* and *High depth* are the UC Captions rows, the three caption depths our pipeline writes for the same span. *Human GT* is the original annotation of the anchor cell.

Every UC Captions row clears the random floor on both datasets, by at least .035 at Low depth (.221 against .186 on YouCook2, .209 against .165 on ViTT), and the margin widens with requested depth, reaching .259 and .240 at High depth. The three depths of a cell share identical timestamps, which makes that rise a caption-content effect rather than a segmentation one. On tIoU every UC Captions row stays above .26, against .078 and .122 for a random caption, so the captions also locate their own span.

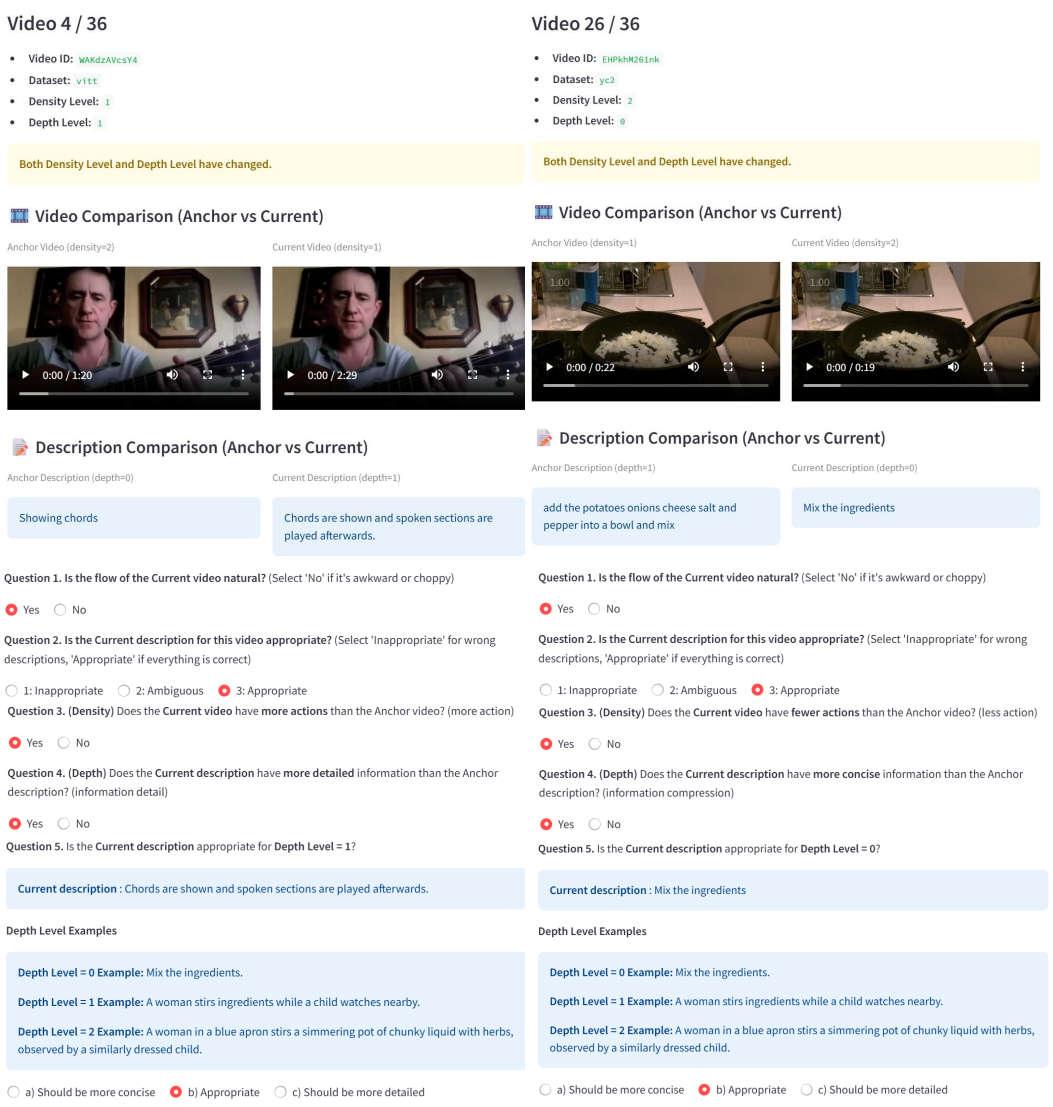


Figure 7: Visualization for our human evaluation protocol.

C Human Evaluation for UC Captions Dataset

C.1 Human Evaluation Protocol

We ran a human evaluation to assess two aspects of the pipeline, the quality of the generated annotations and the perceptual effectiveness of the control axes. We recruited 14 experts from 4 different institutions to keep the pool diverse. Each expert judged 36 randomly sampled video–caption pairs, giving 504 judgments in total.

Figure 7 illustrates the interface presented to experts. Experts were shown video clips and their corresponding captions. Depending on whether the density or depth level of the clip differed from the anchor configuration of the dataset (mid-density, mid-depth for YouCook2; high-density, low-depth for ViTT), they were asked a series of questions designed to evaluate different aspects of our data generation pipeline:

Table 16: Human evaluation results. (a) shows the annotation quality, measured by Temporal Coherence (TC %) and Caption Quality (CQ Score, 1–3). (b) presents the effectiveness of the control mechanisms. Underlined entries indicate the anchor configuration of each dataset.

Dataset	Depth Level	Density Level		
		Low	Mid	High
YouCook2	Low	96.4 / 2.52	92.9 / 2.93	96.4 / 2.82
	Mid	96.4 / 2.93	100 / 2.93	89.3 / 2.56
	High	89.3 / 2.93	100 / 2.75	96.4 / 2.64
ViTT	Low	92.9 / 2.54	82.1 / 2.61	89.3 / 2.45
	Mid	96.4 / 2.89	92.9 / 2.54	80.6 / 2.71
	High	85.7 / 2.93	96.4 / 2.64	100 / 2.75

(a) Annotation Quality.

Dataset	Control Metric	Level	Score
YouCook2	Density Control (%)	Mid → Low	96.4
		Mid → High	88.1
	Depth Quality (1–3)	Low	2.86
		Mid	2.71
		High	2.88
ViTT	Density Control (%)	High → Mid	89.3
		High → Low	97.6
	Depth Quality (1–3)	Low	2.79
		Mid	2.86
		High	2.98

(b) Control Mechanism Effectiveness.

Table 17: Anchor-identity-blinded three-way level matching. Entries are accuracy (%), mean absolute error, and inter-rater agreement (Krippendorff's α).

Axis	YC2	ViTT	Pooled
Depth	98.4/.016/.95	98.4/.016/.95	98.4/.016/.95
Density	92.1/.159/.76	92.1/.143/.77	92.1/.151/.77

- **Temporal Coherence (Q1):** Experts rated whether the flow of the video segment was natural. This assesses the quality of our automated segmentation for varying density levels. A "Yes" indicates a perceptually coherent clip.
- **Caption Quality (Q2):** Experts judged the factual correctness and relevance of the caption on a 3-point scale (1: Inappropriate, 2: Ambiguous, 3: Appropriate). This validates the quality of our caption depth refinement.
- **Control Mechanism Validation (Q3 & Q4):** In cases of density or depth variation, experts performed a comparative judgment against an anchor video or caption. They were asked if the current sample correctly reflected the intended change (*e.g.,* "Does the current video contain more actions?", "Is the current caption more detailed?"). This directly measures the perceptual success of our control dimensions.
- **Absolute Depth Assessment (Q5):** Finally, experts determined if the detail level of a caption was appropriate for its given depth level (Low, Mid, High), with options to classify it as needing to be more concise, appropriate, or more detailed.

C.2 Annotation Quality

We first evaluated the quality of the generated annotations, focusing on the temporal coherence of video segments and the appropriateness of captions. As shown in Table 16(a), the generated video clips maintain a high degree of temporal coherence. For YouCook2, scores were consistently high, averaging over 95%, while for ViTT, they also demonstrated strong coherence, averaging over 90% across all levels. Furthermore, the caption quality scores for all generated levels are comparable to the original human-annotated anchors (*e.g.,* 2.93 for the YouCook2 anchor vs. 2.52–2.93 for others). These results indicate that our pipeline generates high-quality, reliable annotations across the entire 3×3 granularity spectrum.

C.3 Control Mechanism Effectiveness

We then assessed the perceptual effectiveness of our density and depth control mechanisms. The results in Table 16(b) show that our density control is perceptually effective, achieving high accuracy scores for both YouCook2 and ViTT (*e.g.,* 97.6% for the High→Low control on ViTT). The depth

control was also validated, with captions at all levels receiving high appropriateness scores. Both axes are therefore perceptible to a human reader, which is what the dataset needs to support training controllable models.

C.4 Blinded Level-Identification Study

To further corroborate the above findings, we additionally conducted a blinded variant of the study in which the identity of the source anchor is hidden. Each trial presents three shuffled annotations of the same video and asks the expert to assign Low, Mid, and High exactly once, with labeled calibration examples preceding the evaluation and no option identified as the anchor. The same 14 experts judge both datasets, giving 252 judgments per axis, 126 per dataset. As reported in Table 17, experts identify the requested level of both axes far above the 33.3% that a random assignment would reach: depth with 98.4% accuracy and density with 92.1%. The two datasets yield the same number of correct answers (124 of 126 for depth and 116 of 126 for density), while their density errors differ, as the mean absolute errors show ($.159$ and $.143$). This indicates that the requested level can be read from the annotation alone, without any reference to the anchor, providing additional support for the effectiveness of the control mechanisms reported in Table 16.

D Additional Implementation Details for UC-DVC

We integrate UC-DVC into three DVC backbones, Vid2Seq (Yang et al., 2023b), HiCM^2 (Kim et al., 2025), and Sali4Vid (Jeon et al., 2025), and Sali4Vid is the default backbone unless a table states otherwise. Appendix E.3 further attaches it to the DETR-based PDVC. As described in the main paper, we utilized the pretrained T5-Base model with embedding size $h = 768$ for the text encoder and decoder, while the visual encoder was based on the pretrained CLIP ViT-L/14 model. All experiments were implemented using PyTorch, with a batch size of 2 and an initial learning rate of 3e-4. For each density and depth encoder (E_d and E_z), we use 3 encoder transformer layers with the same hidden dimension $h = 768$ as the T5-Base model. We employed the Adam optimizer (Kingma, 2014), setting parameters to (0.9, 0.999), consistent with standard practices. During training, video frames were sampled at a consistent rate of 1 frame per second (fps) and standardized into sequences of 100 frames through padding or subsampling. Training was conducted on a single NVIDIA RTX 5090 GPU for 20 epochs with UC Captions of YouCook2 and ViTT datasets.

Relative Supervision from Target–Reference Pairs. Figure 4 illustrates the two objectives that turn a paired target–reference annotation into training signal, and the main paper states them in Sec. 3.2. **(a) Density-aware loss.** Each training sample carries a target cell and a different reference cell for the same video. Timestamp tokens are reweighted by $w_k = 1/\mathrm{IoU}_i$, where IoU_i averages the best temporal overlap between the target event of the k-th token and the events at each other density level (Section 3.2), and boundaries that move more between density levels receive stronger supervision, while non-timestamp tokens keep unit weight. **(b) Depth-aware loss.** Cross-attention runs from the depth-modulated representation V_z to the encoded target or reference captions, and the saliency token is trained to score a caption at the requested depth as 1 and a caption at another depth as 0. The model therefore learns which linguistic properties should change when the depth request changes, rather than treating the nine cells as unrelated copies.

E Additional Experiments

E.1 Evaluating Generalization Across Different MLLMs and LLMs

To test whether UC Captions depends on the particular generators used to build it, we rebuilt the annotations with different MLLM and LLM pairs, using InternVideo2 (Wang et al., 2024a) and ShareGPT4Video (Chen et al., 2024b) for the MLLM and Llama 3-8B and Llama 3.3-70B for the LLM. In Table 18 every pair clears the source-trained reproduction, with METEOR between 13.68 and 14.02 against 13.57. The stronger pair yields the better corpus, but the gain does not hinge on one generator pair. UC Captions is reproducible with an open generator stack rather than tied to one choice.

Table 18: UC Captions built with different MLLM and LLM generators, evaluated on YouCook2 *val* with Sali4Vid. † is our reproduction (Table 6).

Method	MLLM	LLM	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1
Sali4Vid†	-	-	13.57	76.13	10.41	33.79
Ours	InternVideo2	Llama 3-8B	13.68	79.37	10.55	35.72
		Llama 3.3-70B	13.83	80.73	10.80	36.35
	ShareGPT4Video	Llama 3-8B	13.87	81.36	10.92	36.50
		Llama 3.3-70B	**14.02**	**82.35**	**11.23**	**36.91**

Table 19: Effect of $\mathcal{L}_\theta$ and $\mathcal{L}_s$ on YouCook2 *val* with Sali4Vid.

$\mathcal{L}_\theta$	$\mathcal{L}_s$	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1
✗	✗	11.57	73.96	9.35	34.33
✓	✗	12.20	78.04	10.13	36.67
✗	✓	13.41	80.06	10.63	35.23
✓	✓	**14.02**	**82.35**	**11.23**	**36.91**

Table 20: Full nine-cell UC Captions comparison on the HiCM^2 backbone, a retrieval-augmented DVC model. A single UC-DVC is evaluated against nine HiCM^2 models, each trained on one density–depth cell. Values are baseline / **UC-DVC** (improvement). UC-DVC covers all nine operating points with one shared parameter set, and the comparison reflects model consolidation and control coverage.

		UC Captions (YouCook2)			UC Captions (ViTT)		
Density	**Metric**	**Low Depth**	**Mid Depth**	**High Depth**	**Low Depth**	**Mid Depth**	**High Depth**
Low	$\mathcal{M}$	9.1 / **10.5**(+1.4)	9.2 / **10.3**(+1.1)	8.4 / **9.4**(+1.0)	6.5 / **7.8**(+1.3)	8.5 / **9.5**(+1.0)	9.2 / **10.1**(+0.9)
	F1	31.8 / **39.0**(+7.2)	30.2 / **39.6**(+9.4)	30.7 / **39.5**(+8.8)	44.0 / **51.5**(+7.5)	44.2 / **53.8**(+9.6)	45.1 / **53.7**(+8.6)
Mid	$\mathcal{M}$	10.0 / **11.8**(+1.8)	12.5 / **12.9**(+0.4)	8.8 / **9.7**(+0.9)	6.6 / **8.3**(+1.7)	9.4 / **12.1**(+2.7)	9.2 / **10.0**(+0.8)
	F1	31.7 / **37.6**(+5.9)	32.5 / **37.9**(+5.4)	30.9 / **36.9**(+6.0)	45.5 / **51.2**(+5.7)	45.0 / **51.1**(+6.1)	46.1 / **51.8**(+5.7)
High	$\mathcal{M}$	8.4 / **11.0**(+2.6)	8.5 / **10.6**(+2.1)	7.5 / **9.1**(+1.6)	9.4 / **11.3**(+1.9)	10.8 / **12.7**(+1.9)	9.9 / **11.3**(+1.4)
	F1	27.9 / **30.6**(+2.7)	27.8 / **30.9**(+3.1)	26.4 / **29.9**(+3.5)	44.9 / **46.5**(+1.6)	45.3 / **48.2**(+2.9)	45.0 / **48.3**(+3.3)

E.2 Effect of the Proposed Objectives

Table 19 separates density-aware timestamp weighting from the depth-aware comparative loss. Disabling density weighting does not remove the sequence-generation objective. Instead, all timestamp weights are set to $w_k = 1$, following the conventional formulation (Kim et al., 2025). Density weighting provides the larger F1 gain, while the depth-aware comparative loss provides the larger improvement in captioning metrics. Combining both components achieves the best result across all reported metrics.

E.3 Generalization Across Different DVC Methods

As shown in Table 6 of the main paper, integrating our UC-DVC framework into three different DVC backbones—Vid2Seq, HiCM^2, and Sali4Vid—yields consistent improvements on both the standard YouCook2 and ViTT benchmarks, confirming that our approach generalizes across architectures rather than depending on a specific backbone. Moreover, we further validate this generalizability under the user-controllable setting by adopting HiCM^2 as the backbone.

Table 20 repeats the coverage comparison on HiCM^2. One shared parameter set answers all nine requests and stays above the HiCM^2 model trained on that cell in every cell on both YouCook2 and ViTT. The consolidation therefore carries over to a retrieval-augmented architecture.

We further conduct detailed ablation studies on the HiCM^2 backbone, mirroring the Sali4Vid studies above for a side-by-side reading. Table 21 varies the MLLM and the LLM used to build UC Captions. Every combination improves on the HiCM^2 baseline and the spread across the four combinations stays within 0.30 METEOR, which indicates that the gain does not depend on a particular generator pair. Table 22 removes each learning objective in turn. The density-aware loss alone recovers most of the localization gain (37.66 F1 against 37.90 for the full model) and the depth-aware loss alone recovers most of the captioning gain (12.31 METEOR against 12.92), the same division of labor observed on Sali4Vid in Table 19. Tables 23 and 24 repeat the two component studies of the main paper. MaxDiv and TKG each help on the data side and combine best, and the density and depth modules each help on the modeling side and combine best, where removing both controls drops the model to 8.38 METEOR. Every study reproduces on a retrieval-augmented

Table 21: Experimental results on the YouCook2 *val* set using different MLLMs and LLMs, with the HiCM^2 backbone. † denotes our reproduction, following Table 6 of the main paper.

Method	MLLM	LLM	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1
HiCM^{2†}	-	-	12.52	68.92	10.71	32.54
Ours	InternVideo2	Llama 3-8B	12.62	72.58	10.92	35.87
		Llama 3.3-70B	12.75	74.15	11.28	36.94
	ShareGPT4Video	Llama 3-8B	12.79	74.82	11.45	37.18
		Llama 3.3-70B	**12.92**	**75.97**	**11.91**	**37.90**

Table 22: Effect of proposed losses $\mathcal{L}_\theta$ and $\mathcal{L}_s$ on YouCook2 *val* set, with the HiCM^2 backbone.

$\mathcal{L}_\theta$	$\mathcal{L}_s$	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1
✗	✗	10.47	67.58	10.03	35.32
✓	✗	11.10	71.66	10.81	37.66
✗	✓	12.31	73.68	11.31	36.22
✓	✓	**12.92**	**75.97**	**11.91**	**37.90**

Table 23: Effect of proposed component in data generation on YouCook2 *val* set, with the HiCM^2 backbone. † denotes our reproduction, following Table 6 of the main paper.

Data Type	MaxDiv	TKG	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1
Original†	-	-	12.52	68.92	10.71	32.54
UC Captions	✗	✗	12.82	72.45	11.25	34.87
	✓	✗	12.84	73.97	10.86	34.72
	✗	✓	12.88	74.39	11.18	37.25
	✓	✓	**12.92**	**75.97**	**11.91**	**37.90**

Table 24: Effect of proposed component on YouCook2 *val* set, with the HiCM^2 backbone.

Density	Depth	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	F1
✗	✗	8.38	43.72	8.40	31.00
✓	✗	10.75	55.99	10.37	36.42
✗	✓	12.21	70.42	10.73	34.95
✓	✓	**12.92**	**75.97**	**11.91**	**37.90**

Table 25: Transfer to the DETR-based PDVC on YouCook2 *val*.

Method	$\mathcal{M}$	F1
PDVC	5.56	26.81
PDVC + UC-DVC	**6.01**	**28.77**

Table 26: Design choices for the depth pathway with Sali4Vid on YouCook2 *val*.

Condition	$\mathcal{M}$	F1
Concat + Self-Attn	13.87	36.20
Cross-Attn	**14.02**	**36.91**
CLS-token	13.41	36.12
Saliency-token	**14.02**	**36.91**

Table 27: MLLM-based DVC and Sali4Vid with and without UC-DVC on YouCook2 *val*.

Method	$\mathcal{C}$	$\mathcal{S}$	F1
TimeChat (Ren et al., 2024)	11.0	3.4	19.5
VTG-LLM (Guo et al., 2025a)	13.4	3.6	20.6
TRACE (Guo et al., 2025b)	35.5	6.7	31.8
TimeExpert (Yang et al., 2025)	39.0	7.2	33.5
Sali4Vid (Jeon et al., 2025)	75.80	10.28	33.61
Sali4Vid + UC-DVC	**82.35**	**11.23**	**36.91**

backbone, which supports the claim that the controllable behavior belongs to UC-DVC rather than to one architecture.

We finally transfer the framework to a structurally different family. PDVC localizes events by parallel decoding rather than by generating a sequence, and Table 25 shows that adding UC-DVC raises METEOR from 5.56 to 6.01 and F1 from 26.81 to 28.77. The absolute scores sit far below the sequence-based backbones because PDVC is a much older and smaller model, and the result shows that the control modules attach to a DETR decoder and still improve it.

E.4 Design Choices

Table 26 compares the architectural alternatives under a common data and evaluation protocol. Cross-attention outperforms concatenation with self-attention, matching the asymmetric roles of the depth query and the caption evidence, and the saliency token outperforms a generic CLS aggregator.

E.5 Comparison with MLLM Models

We compare UC-DVC with recent MLLM-based dense video captioning approaches on the YouCook2 *val* set in Table 27 (extending Table 2 of the main paper with UC-DVC). Overall, existing MLLM-based methods remain substantially behind on core DVC metrics, highlighting that strong general video-language generation does not necessarily translate into temporally grounded, event-level captions under the standard evaluation. Even Sali4Vid trained only on the original annotation leads the strongest MLLM, TimeExpert, by 36.8 CIDEr and 3.1 SODA$_c$, while their

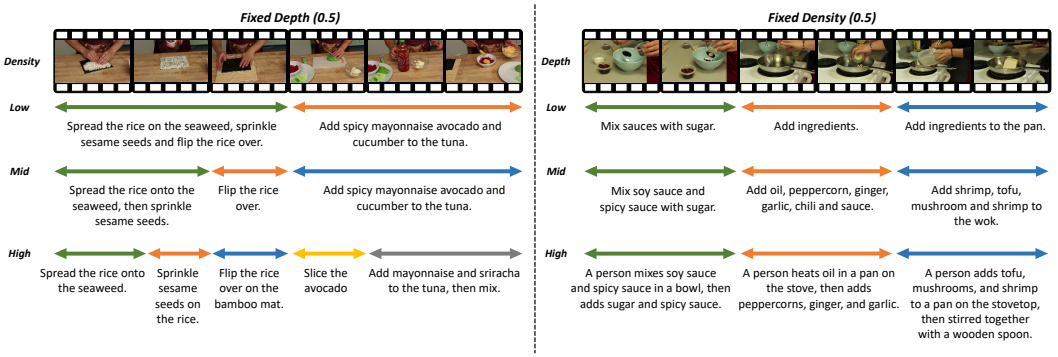


Figure 8: Visualization results of density and depth variations of the proposed UC-DVC on YouCook2 *val* set.

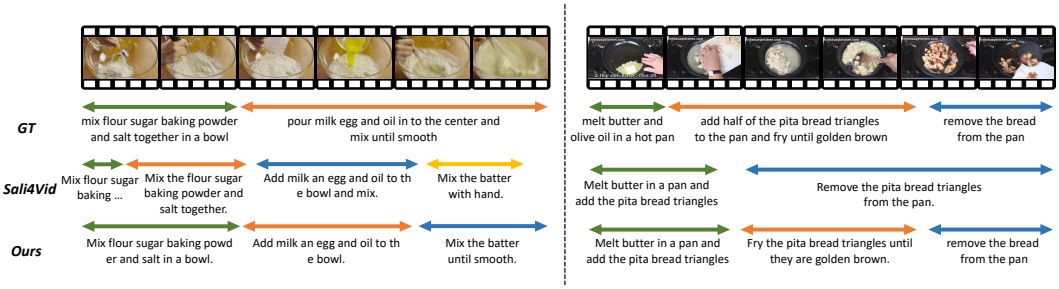


Figure 9: Visualization results for YouCook2 *val* set. We compare our method with Sali4Vid.

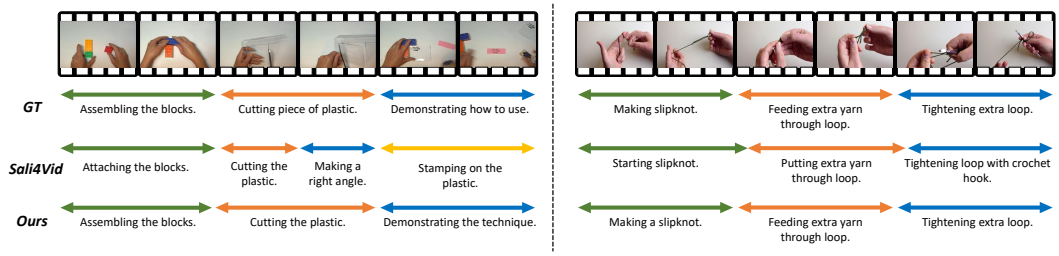


Figure 10: Visualization results for ViTT *test* set. We compare our method with Sali4Vid.

F1 is close (33.61 vs. 33.5), so the gap lies mainly in describing each localized event rather than in finding it. Adding UC-DVC to Sali4Vid raises all three metrics further (82.35, 11.23, 36.91), indicating that our framework better captures fine-grained event structure within the DVC setting.

F Additional Visualization Results

Figure 8 visualizes density and depth control on YouCook2. With depth fixed, increasing density progressively divides the video into more localized events while retaining a similar level of caption detail. With density fixed, increasing depth enriches each caption with additional actions, objects, and contextual information without substantially changing the event structure.

F.1 Comparison with State-of-The-Art Framework

Figure 9 and Figure 10 provide qualitative comparisons among ground-truth annotations, predictions from Sali4Vid, and our proposed UC-DVC. While both models predict meaningful event boundaries and captions, UC-DVC produces more accurate and contextually rich captions and boundaries compared to Sali4Vid. These examples suggest that our proposed loss functions effectively allow the model to leverage additional semantic information from various density and

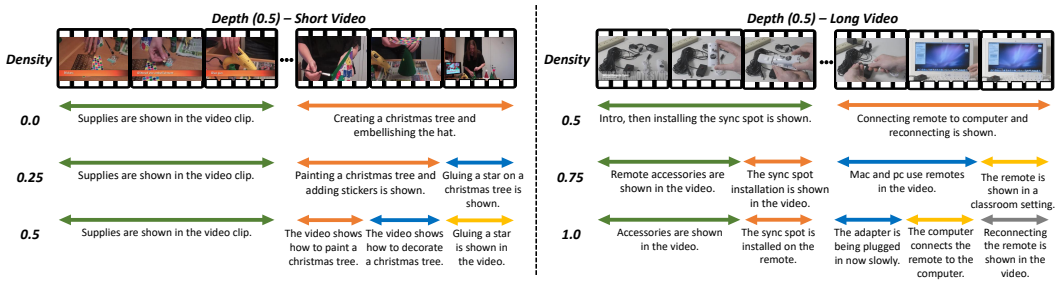


Figure 11: Visualization results for intermediate density levels on the ViTT *test* set. The figure shows outputs for density values of 0.25 and 0.75.

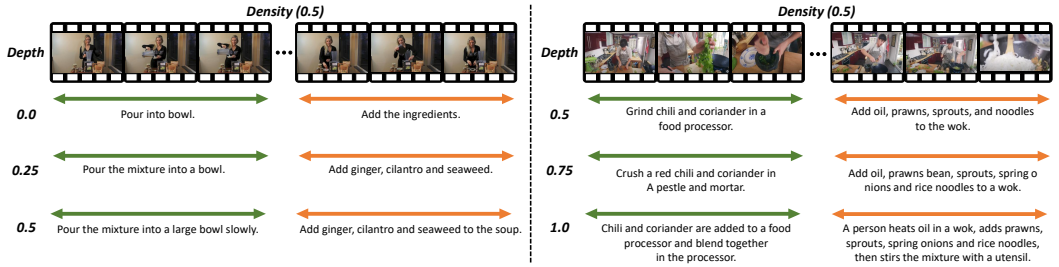


Figure 12: Visualization results for intermediate depth levels on the YouCook2 *val* set. The figure shows outputs for depth values of 0.25 and 0.75.

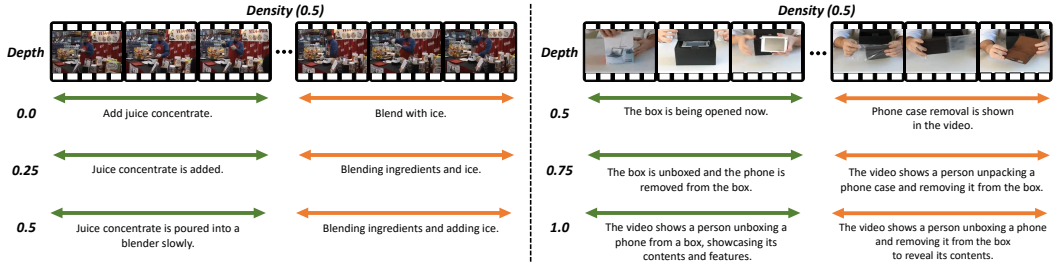


Figure 13: Visualization results for intermediate depth levels on the ViTT *test* set. The figure shows outputs for depth values of 0.25 and 0.75.

depth annotations during training.

F.2 Fine-grained User Control

Figures 11, 12, and 13 extend the intermediate-value visualization of the main paper to the density axis on ViTT and to the depth axis on both datasets, going beyond the discrete Low (0.0), Mid (0.5), and High (1.0) levels. For density control, Figure 11 shows that the model generates an intermediate number of events for a density value of 0.25, producing more events than the Low level and fewer than the Mid level. For depth control, Figures 12 and 13 illustrate that, at a depth value of 0.75, the model generates captions that are more descriptive than Mid-level sentences but still less verbose than High-level narratives. Taken together, these results confirm that UC-DVC responds in the requested order at values it never saw during training, on both datasets and on both axes.

F.3 Visualization of General Domain Cases

Figure 14 presents qualitative results of UC-DVC trained on ViTT on general-domain videos from outside both datasets. The figure demonstrates the capability of the model to accurately localize and

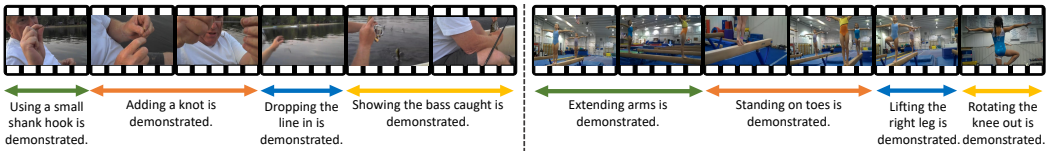


Figure 14: Visualization results for General Domain Videos.

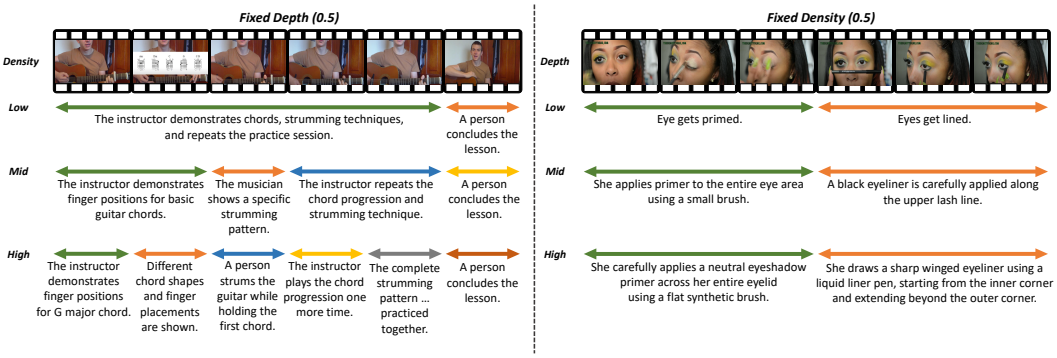


Figure 15: Visualization results of density and depth variations of the proposed UC-DVC on ViTT *test* set.

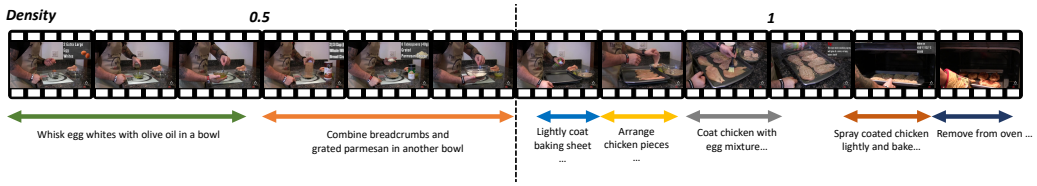


Figure 16: Visualization results for streaming scenario on the YouCook2 *val* set.

describe diverse activities, ranging from intricate procedural steps in fishing (*e.g.,* "Adding a knot", "Dropping the line in") to precise body movements in physical exercises (*e.g.,* "Standing on toes", "Rotating the knee out"). These examples come from outside the two anchor domains and indicate that the same control behavior appears there as well.

F.4 Visualization on the ViTT Test Set

Figure 15 shows visualization results of the proposed UC-DVC under varying density and depth conditions on the general-domain ViTT *test* set, confirming that our approach generalizes beyond the cooking-focused YouCook2 domain.

F.5 Visualization of Streaming Scenario

Figure 16 validates the adaptability of UC-DVC in a real-time streaming scenario where control requirements change dynamically. In this visualization, we shift the density input from 0.5 (Mid) to 1.0 (High) during the continuous video stream. The model demonstrates immediate responsiveness, seamlessly transitioning from generating broader event descriptions to capturing fine-grained actions as the density signal increases. Event granularity therefore tracks the request while the stream continues, which is the behavior an interactive setting needs.

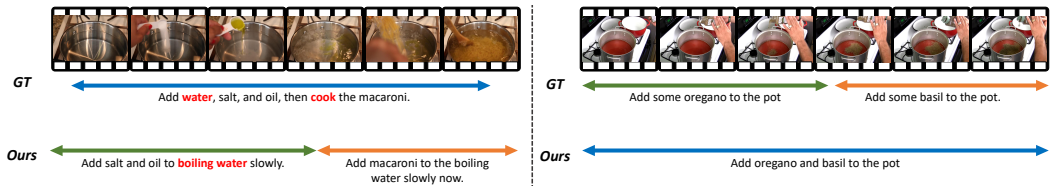


Figure 17: Visualization results for failure cases on YouCook2 *val* set.

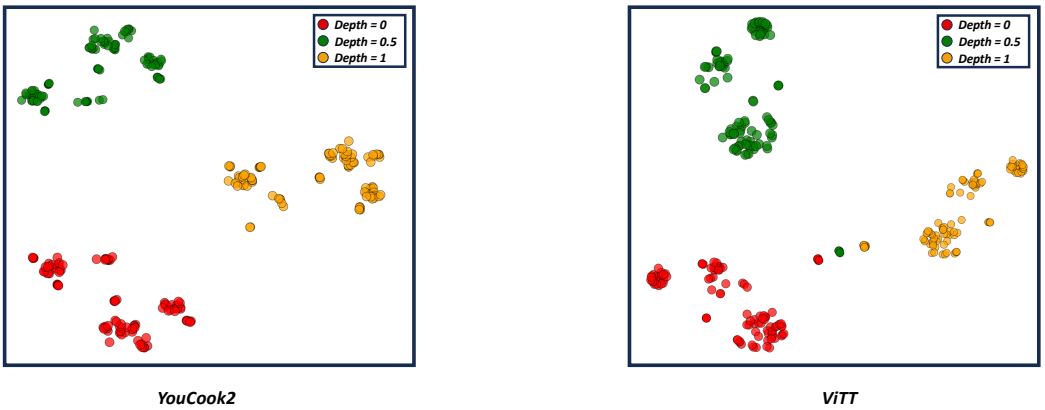


Figure 18: t-SNE visualization of depth encoder features for depth values.

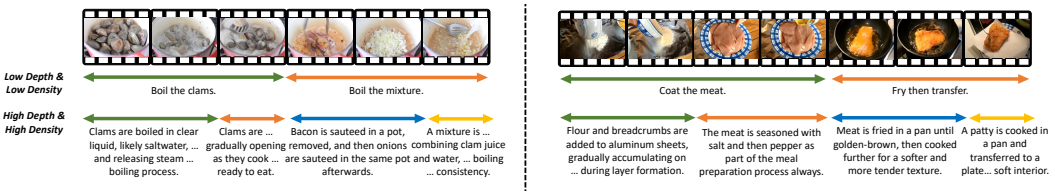


Figure 19: Low-Density/Low-Depth and High-Density/High-Depth annotations in UC Captions.

F.6 Visualization of Failure Cases

Figure 17 analyzes representative error modes encountered by our framework. The left panel illustrates an instance of *over-segmentation* combined with *temporal state hallucination*. While the ground truth defines a single continuous event, the model incorrectly fragments it into two segments. Furthermore, the visual prominence of "boiling water" in the latter frames retroactively influences the first segment, causing the model to hallucinate that the initial water is already "boiling" before the heating process completes. The right panel depicts an *under-segmentation* error where two temporally distinct atomic actions (adding oregano and adding basil) are merged into a single coarse caption due to their high visual similarity. These cases highlight the challenges in maintaining physical state consistency across segmented events and distinguishing visually analogous actions in rapid succession.

F.7 Impact of Control Parameters on Latent Representations

To further validate that our control parameters effectively influence the learned representations, we visualize the latent representations produced by the depth encoder using t-SNE dimensionality reduction in Figure 18. We extract features corresponding to depth values of 0.0, 0.5, and 1.0 from the depth encoder and project them into a two-dimensional space with t-SNE. The visualization reveals clear clustering patterns, with features from the same depth level tightly grouped together, while features from different depth levels are distinctly separated in the latent space. The depth encoder therefore carries depth-dependent representations, which is where the control input takes effect. Such separation in the feature space enables the model to generate captions with varying levels of descriptive detail according to user demand.

F.8 Visualization of Contrasting Granularity Settings

Figure 19 contrasts Low-Density/Low-Depth and High-Density/High-Depth annotations. The former summarizes broader temporal regions with concise captions, whereas the latter divides the video into finer events and expresses more descriptive information for each event. These examples qualitatively illustrate that the generation pipeline produces visibly distinct annotations at opposite ends of the control grid. They are provided as representative cases rather than exhaustive evidence of factual correctness.

F.9 Visualization of the Textual Knowledge Graph

As shown in Figure 20, we visualize the textual knowledge graphs (TKG) built in our UC Captions pipeline from the original caption and the MLLM caption. The graphs show which facts each caption level draws on.

F.10 Evaluation Prompts for LLM-Based Semantic Assessment

To ensure transparency and reproducibility of the LLM-based semantic evaluation, we provide the exact prompts used in our assessment protocol. As shown in Figure 21 and Figure 22, the evaluation consists of two complementary prompts designed to measure the semantic depth of captions.

G Source Code & Benchmark

To support reproducibility during the review process, we provide the source code for the data generation pipeline in the supplementary material, which lets a reader assess how the benchmark was constructed. Upon acceptance, we will publicly release the full source code and the benchmark with documentation, for the community to access and use for research.

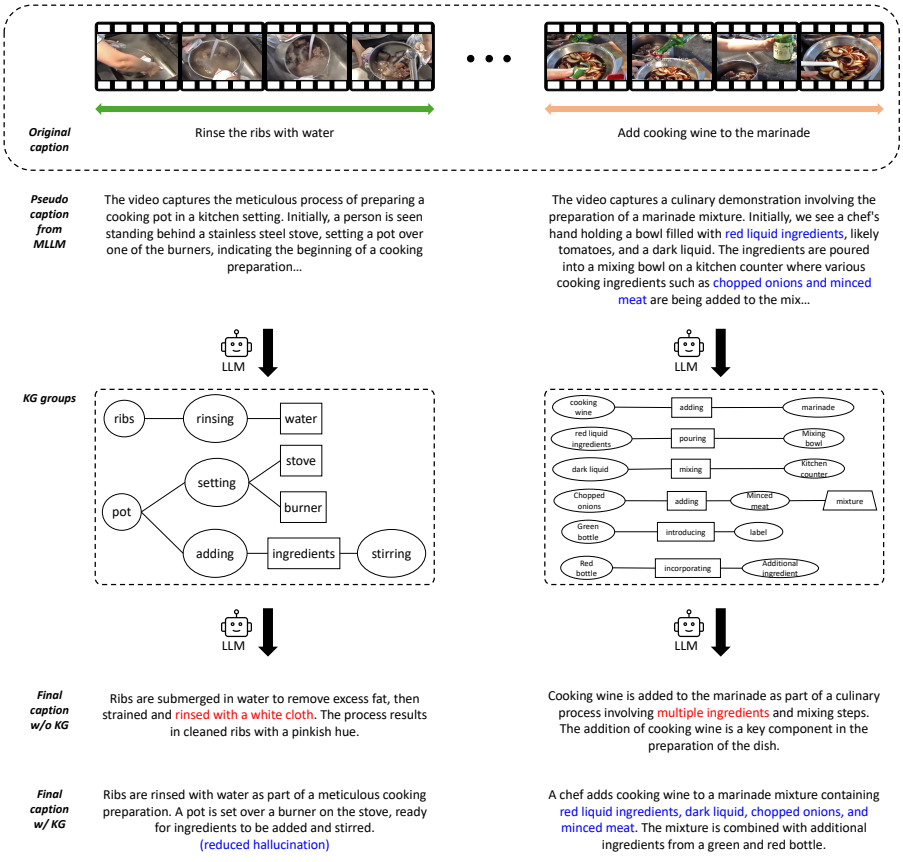


Figure 20: Visualization Results of Knowledge Graph.

Scoring Prompt (YouCook2)

System Instruction
You are an expert evaluator for COOKING video captions. Rate 'Semantic Depth' (0.0-1.0).

Definition
Depth : Density of specific cooking details (Agent, Action, Ingredient, Tool, Context).

Definition of Semantic Depth
'Semantic Depth' is NOT about simple word count. It is defined by the density of specific cooking details.
Evaluate based on the richness of details regarding:
1. [Agent]: Who (e.g., "a chef" > "a person")
2. [Action]: How (e.g., "finely chops" > "cuts")
3. [Ingredient/Object]: What (e.g., "fresh basil" > "herbs")
4. [Tool/Context]: Where/With what (e.g., "in a non-stick pan" > "in a pan")

Task
Score these three captions based on the definition above.

Rubric
0.0: "A person is shown." (Vague) / 0.5: "A woman cuts vegetables." (Basic action) / 1.0: "A woman in a red shirt slices carrots on a wooden board." (Detailed)

Input Captions
A: "{A Caption} / B: "{B Caption}“ / C: "{C Caption}"

Output Format
Respond JSON: { "A": 0.x, "B": 0.x, "C": 0.x }

Figure 21: Scoring Prompt for UC Captions (YouCook2).

Ranking Prompt (YouCook2)

System Instruction
You are an expert evaluator for COOKING video captions. Your task is to rank the following three captions describing the same event from LEAST Semantic Depth to MOST Semantic Depth.

Definition
Depth : Density of specific cooking details (Agent, Action, Ingredient, Tool, Context).

Definition of Semantic Depth
'Semantic Depth' is NOT about simple word count. It is defined by the density of specific cooking details.
Evaluate based on the richness of details regarding:
1. [Agent]: Who (e.g., "a chef" > "a person")
2. [Action]: How (e.g., "finely chops" > "cuts")
3. [Ingredient/Object]: What (e.g., "fresh basil" > "herbs")
4. [Tool/Context]: Where/With what (e.g., "in a non-stick pan" > "in a pan")

Task
Rank these three captions based on the definition above.

Input Captions
A: "{A Caption}" / B: "{B Caption}“ / C: "{C Caption}"

Respond with a JSON object containing the list of letters in order (Least -> Middle -> Most)

Output Format
{"ranking_order": ["Letter_Least", "Letter_Middle", "Letter_Most"] }

Figure 22: Ranking Prompt for UC Captions (YouCook2).