Causal-VLM: Dense Causal Captioning in Videos

Asmar Nadeem[1*] **Mahrukh Awan**[1] **Muhammad Awais**[1] **Robert Dawes**[2]
Adrian Hilton[1] **Armin Mustafa**[1]

[1]**CVSSP, University of Surrey, UK** [2]**BBC Research and Development, UK**

Abstract

Dense video captioning describes multiple events in long videos. However, each event is predicted independently, ignoring the causal relationships with the other events in the video. These causal relationships denote observed prerequisite dependencies and are important for human-like descriptions of long videos. Our experiments show that existing vision-language models fail at predicting causal relationships in long-form videos. We address this through two contributions. First, we create a benchmark dataset with 21.6K videos and 85K events by using a large language model to generate annotations from dense captions in YouCook2 and ActivityNet. We validate this benchmark dataset through multimodal alignment, vision-language models, and human judgment. Second, we propose a novel architecture for dense causal captioning that not only generates captions but also learns causal relationships. This novel architecture not only predicts causality effectively (0.73 F1 ActivityNet, 0.64 F1 YouCook2) but also improves the vision-language model's core capabilities: caption quality increases by +6.3 CIDEr on ActivityNet and +7.3 on YouCook2, while temporal grounding becomes accurate as causal predictions force precise event localization. Project page: `https://causalvlm.github.io/`.

1 Introduction

Understanding causal relationships in videos requires predictions beyond temporal sequences to identify why events occur and how they relate. In long-form videos, events form complex causal structures including parallel chains [1, 2], where multiple independent events must occur to enable a subsequent event. Consider making fried chicken (Figure 1): coating the chicken (E6) requires both seasoning the chicken (E4) and separately seasoning flour (E5). This differs fundamentally from temporal ordering; both E4 and E5 must complete before E6 can begin. This enables instructional video systems to identify critical prerequisite steps for effective tutorials, robotics systems to plan sequential tasks and diagnose execution failures, and accessibility tools to provide explanatory descriptions that convey why events occur rather than merely describing what happens.

Recent work addresses causality in video captioning through short-clip single causal caption [3], question-answering [4, 5], or synthetic environments [6, 7]. [8, 9] tackle multi-event causal discovery by identifying which relationships exist among a given set of pre-segmented events, but do not address dense causal caption generation in long-form videos. Current dense video captioning models [10–15] generate event descriptions but cannot infer event causal relationships (coating requires prior seasoning). We introduce *dense causal captioning*: given a raw video, we generate (1) N temporally grounded events with captions, and (2) a causal adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ where $A_{ij} = 1$ indicates event i causes event j, an observed prerequisite dependency (Section 3.2).

A key obstacle is the lack of datasets with dense causal-caption annotations for long-form videos. Existing dense captioning datasets [16, 17] provide event timestamps and captions but lack causal labels. CTN dataset [3] contributes causal labels but is limited to short clips with single cause-effect pairs. To address this, we introduce a new **Dense Causal Captions (DCC)** benchmark dataset. We use Llama-3.3-70B [18] with structured prompting to generate causal annotations from dense captions. Annotations are rigorously validated through three established methods as in [3, 8, 9, 12]: multimodal alignment (EMScore [19]), vision-language model (VLM) evaluation (InternVL3.5-241B [13]), and human validation. DCC enables prerequisite-style structural causal *prediction* on observational video

*Corresponding author: `asmar.nadeem@surrey.ac.uk`

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

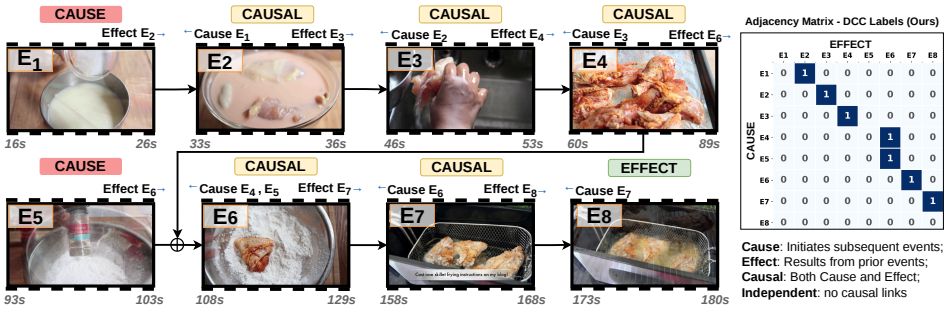

Figure 1: Causal structure in long-form videos. Making fried chicken demonstrates complex causal dependencies across eight events: first chain (E1→E2→E3→E4), parallel preparation (E5), and then coating (E6) requires both E4 and E5, and final steps (E6→E7→E8). The adjacency matrix (right), provided by our Dense Causal Captions (DCC) benchmark dataset, reveals only 7 causal pairs among 28 possible, highlighting the sparse nature of causal relationships in long-form videos.

in the sense [3, 4, 8] and captures multi-event causal relationships along with generating dense captions in long-form videos. The causal annotations are packaged in the form of an adjacency matrix (see Figure 1).

Next, we evaluate the state-of-the-art VLMs: PLM-3B [12], InternVL-3.5-4B [13], Qwen3-VL-8B [20], and InternVL3.5-8B [13], on the DCC benchmark dataset in zero-shot (ZS) and fine-tuned (FT) settings, alongside closed-source models Gemini 2.5 Pro [21] and GPT-5 [22] in zero-shot. Even after fine-tuning, open-source models fail at predicting causal relationships. To address this, we propose the **Causal-VLM**, a two-stage multimodal framework that, using video-only input, can generate captions, event timestamps and predict causal relationships over dense multi-event sequences in videos. Our key insight is that visual features alone cannot capture semantic causal prerequisites; in Figure 1 fried chicken example, "season chicken" and "season flour" appear independent yet both are required for coating. Causal-VLM trains a dense captioning model with curriculum learning [23], generating dense, temporally-grounded event captions and then, adds a multimodal causal head predicting pairwise relationships through cross-modal fusion, contextual encoding, and techniques for extreme class imbalance [24, 25] (to address sparsity, Figure 1).

To summarise, existing dense captioning methods and VLMs ignore causal dependencies in long-form videos. There are no benchmark datasets to address dense causal captioning. We make the following contributions:

- **DCC benchmark**: First benchmark that pairs dense captions and timestamps with a full pairwise causal graph over raw long-form video, with 21.6K videos and 85K events, validated through multimodal alignment, VLM evaluation, and a human study.
- **Causal-VLM**: Two-stage architecture integrating a VLM with a causal head that fuses visual and text features.
- **Results**: Causal-VLM outperforms existing VLMs and transfers to prior CTN [3] and NExT-QA [4] datasets. 0.73/0.64 F1 on causal prediction and +6.3/+7.3 CIDEr on caption quality on ActivityNet/YouCook2, and +0.24/+0.27 F1 over passing Stage-1 captions to Llama-3.3-70B [18].

2 Related Work

Causal Video Benchmarks. Benchmarks for video causal learning have evolved from synthetic environments to real-world videos, yet most remain limited in scope.

Table 1: Comparison of video understanding tasks and causal discovery methods.

Method	Input	Task Type	Output Format	#Vids
Dense Video Captioning (DVC)				
Vid2Seq [14]	Raw	DVC	Cap+Time	–
StreamDVC [15]	Raw	DVC	Cap+Time	–
Video QA / Causal Learning				
NExT-QA [4]	Raw	VQA	Answer	5.4K
C-VidQA [26]	Raw	VQA	Answer	10.7K
CausalVQA [6]	Raw	VQA	Answer	1.6K
Causal Discovery & Narrative				
MECD/+ [8]	Clips+Txt	Discovery	Binary ($N-1$)	1.1K
NarrativeB. [3]	Clip	Narrative	1 Caption	11.9K
C-VLM (Ours)	**Raw**	**Dense+Causal**	**Cap+Time+Mat.**	**21.6K**

CLEVRER [5] provides controlled scenarios with procedurally generated annotations but uses simplified physics that limits generalization. Video question-answering benchmarks like NExT-QA [4] and Causal-VidQA [26] include causal questions but conflate causal learning with commonsense retrieval. More recent work has extended evaluation to counterfactual and hypothetical questions: CausalVQA [6] evaluates physical causal understanding in real-world scenarios, while CausalStep [7] enforces stepwise learning protocols. CLEVRER-Humans [27] collects

human-labelled causal event graphs for synthetic physics videos, Black Swan [28] evaluates abductive and defeasible reasoning about unexpected events, CausalChaos! [29] poses causal questions over longer chains in animated scenes, and Causal2Needles [30] tests joint retrieval of cause and effect events in long videos. However, these benchmarks focus on question-answering rather than generating dense causal captions. Causal captioning benchmarks address generation but remain limited. CTN [3] introduces causal-temporal narrative captions using LLM-based generation, but focuses on short clips (5-20 seconds) with single cause-effect pairs.

MECD [8] and MECD+ [9] address multi-event causal discovery but differ fundamentally in task formulation and scope. Given N pre-segmented events with video clips and textual captions, these predict which events $\{e_1, \ldots, e_{N-1}\}$ cause the final event e_N (N-1 binary decisions). In contrast, our task **dense causal captioning** requires: (1) generating dense captions from raw video, (2) localizing event timestamps, and (3) predicting causal relationships for all $\binom{N}{2}$ ordered pairs, enabling full causal relationship discovery. We provide a detailed analysis in the Appendix E. As summarized in Table 1, existing methods are limited to VQA, single-clip narratives, or causal discovery on pre-processed inputs.

VLMs for Video Understanding. VLMs [31–36] have achieved strong performance on video understanding through large-scale pretraining and architectural contributions. Early models like BLIP-2 [10] introduced Q-Former for efficient visual-linguistic alignment, while Flamingo [37] demonstrated few-shot learning through gated cross-attention. Recent models employ simpler architectures: LLaVA [11] uses linear projection, while LLaVA-1.5 [38] extends this with MLP connectors and higher resolution. Dense video captioning models [14, 15] generate temporally-grounded event descriptions. Models like PLM [12], InternVL [13, 39], Qwen3-VL-8B [20], and Video-LLaVA [40] demonstrate capabilities on standard benchmarks. However, our experiments show that these models struggle with predicting causal relationships between events in long-form videos. This motivates our Causal-VLM architecture combining visual and semantic features for causal relationship prediction between events along with dense caption generation.

Procedure Planning vs. Causal Learning. Procedure planning [41, 42] predicts the *next step* in an instructional sequence given a goal. Our work is fundamentally distinct: we predict *why* each event depends on prior ones, identifying which specific prior events are observed prerequisites for each subsequent event. Parallel causal chains, where event E6 requires *both* E4 and E5 simultaneously, cannot be captured by any sequential step-prediction model. We annotate what occurs in the video, grounded in the observed causal structure, not a generic procedural ontology.

LLM-Supervised Dataset Construction. LLMs enable scalable dataset annotation [3, 8, 10, 11], but generated annotations require validation to prevent errors and biases. Following best practices [3, 43, 44], we employ a three-tier validation pipeline and a blind full-graph human audit (see Section 3.2.2).

Temporal vs. Causal Understanding. Temporal reasoning [45–47] identifies event sequences and ordering; causal understanding [1, 2, 48, 49] determines that one event produces another. Our work demonstrates that models trained only on temporal data fail to make causal predictions.

3 Method

Our approach includes a new benchmark for dense causal captioning (DCC) and a multimodal architecture that learns this task. We first evaluate the existing VLMs, then define the DCC annotations and their construction, and finally, explain Causal-VLM.

3.1 Vision-Language Models Causality Prediction

We first investigate whether the state-of-the-art VLMs can predict causal relationships through zero-shot prompting. We evaluate PLM-3B [12], InternVL-3.5-4B [13], Qwen3-VL-8B [20], InternVL3.5-8B [13], and closed-source models Gemini-2.5- Pro [21] and GPT-5 [22] by providing them with videos and instructing them to output dense video captions and causal relationships. Table 2 presents results across open-source (ZS and FT) and closed-source (ZS) settings. This reveals a fundamental limitation: existing models describe individual actions but cannot reason about causal dependencies. We hypothesize that this occurs because standard vision-language pretraining lacks explicit supervision for causal relationships, and transformer architectures lack mechanisms for distinguishing correlation from causation. This motivates our DCC benchmark dataset for explicit supervision of causal relationships (Section 3.2) and the Causal-VLM specialized architecture (Section 3.3).

In zero-shot evaluation, PLM-3B produces a valid output structure (JSON) in 33% of cases, with 78% of failures in predicting the causal relationships. For instance, in the fried chicken video (Figure 1), the model predicts causality between "combine ingredients" (E1) and "coat chicken" (E6), missing

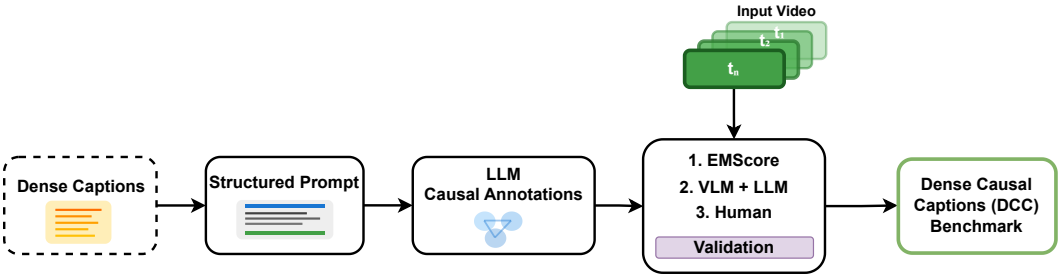

Figure 2: Dense Causal Captions (DCC) annotation generation and validation pipeline.

the required intermediate steps (brining, rinsing, seasoning) and the parallel preparation of seasoned flour (E5). InternVL-3.5-4B achieves 31% valid output with 71% failure rate.
Fine-tuning both models on the new DCC benchmark improves format adherence to 94% for PLM-3B and 91% for InternVL-3.5-4B, but failure rates for causal relationships remain high at 65% and 61%, respectively. The models produce correctly formatted outputs with incorrect causal relationships. Scaling model size does not resolve the failure: Qwen3-VL-8B (FT) and InternVL3.5-8B (FT) improve on the dense captioning task, but the performance degrades on causal prediction. Closed-source models (Gemini-2.5-Pro, GPT-5) perform worst zero-shot (F1 of 0.16/0.12, see Section 4.2).

3.2 Dense Causal Captions (DCC) Benchmark Dataset

For DCC benchmark dataset, each video is annotated with N timestamped event captions and a directed adjacency matrix $\mathbf{A} \in \{0,1\}^{N\times N}$. The main component, causal edge, is defined below:
Causal edge. Let events be indexed by onset time. An edge (i, j) with $i < j$, written $A_{ij} = 1$, represents an observed prerequisite dependency: within the video as filmed, event i is an initiating event that enables event j. It does not assert universal, counterfactual or interventional necessity. Transitive closures are not annotated: for $i \to k \to j$, the edge (i, j) is included only where i directly enables j independent of k. For temporally overlapping events, edge direction follows the onset order. We generate the causal annotations from human labelled dense captions using an LLM (Section 3.2.1) and validate them against the videos (Figure 2) along with blind full-graph audit in Section 3.2.2.

3.2.1 DCC Benchmark Dataset Generation:

We employ an LLM, Llama-3.3-70B [18], similar to [3, 8], to generate causal annotations from dense captions in ActivityNet [16] and YouCook2 [17]. We employ structured prompting that enforces hard constraints while preserving reasoning capability. The prompt (see Appendix C.1) provides four components: (i) explicit event type definitions (*cause*: initiates subsequent events; *effect*: results from prior events; *causal*: both cause and effect; *independent*: no causal links), (ii) temporal consistency requiring causes precede effects in time, (iii) cardinality constraints requiring exactly N events matching the input, and (iv) output JSON format requirements.

3.2.2 Dataset Quality Validation:

LLM outputs require multi-tier validation. We employ three complementary validation methods and a separate full-graph human audit:
1. Multimodal Alignment (EMScore). For each causal pair $(\mathcal{C}, \mathcal{E})$, we merge video segments and compute EMScore [19] ($\theta = 0.2$, following NarrativeBridge [3]). This validates temporal coherence: 94.3% of pairs exceed threshold on ActivityNet, 91.8% on YouCook2.
2. VLM-Based Content Verification. Following [12], we use InternVL3.5-241B [13] to generate captions from merged video clips, then evaluate four dimensions via Llama-3.3-70B [18]: plausibility, visual evidence, strength, and transition. Crucially, Llama judges causal validity from what the VLM *sees* in the merged clips, along with the original human-annotated captions (see Appendix C.2). Mean scores: 0.628 ActivityNet, 0.632 YouCook2. *Critically*, here VLMs caption individual proposed pairs with explicit visual grounding, not full causal graphs over $\binom{N}{2}$ pairs where failure occurs.
3. Human Validation. Following [3, 43, 44], we sample 2,000 videos (1200 ActivityNet, 800 YouCook2) containing 9,574 causal pairs, stratified to oversample videos with 5+ events where annotation risk is highest). Three annotators view merged clips and rate causality (1=definitely not, 5=definitely yes), achieving an intraclass correlation coefficient (ICC) of 0.85 (95%, CI: 0.81-0.89). Results: 98.7% score ≥ 3; 91.7% score ≥ 4. Annotators can see each proposed pair; human and LLM labels agree with $\kappa = 0.76$, where κ is Cohen's coefficient of inter-rater agreement [50].
4. Blind Full-Graph Human Audit. We also do a full graph human audit so that annotators watch the videos and label every eligible event pair in 200 videos (100 ActivityNet, 100 YouCook2),

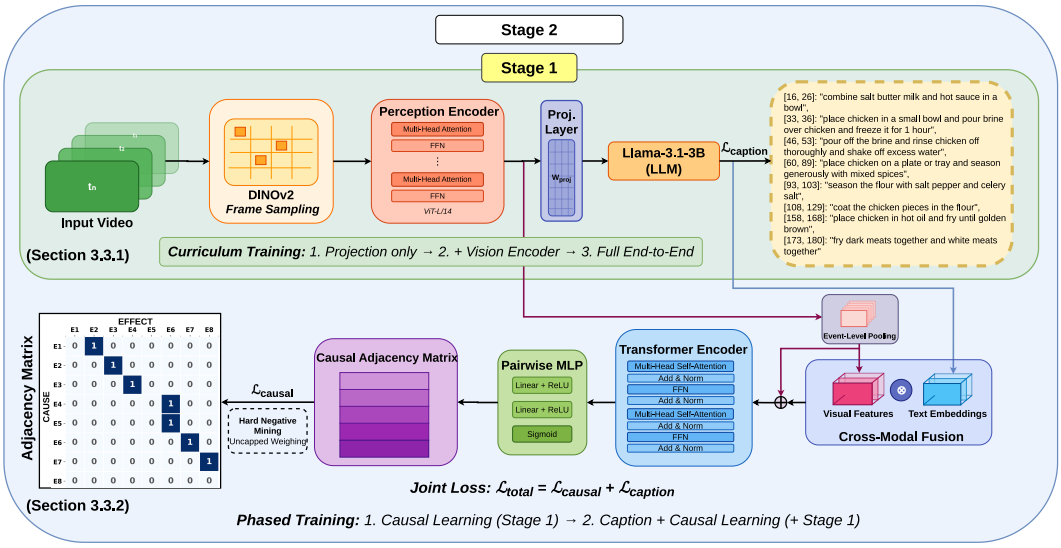

Figure 3: Causal-VLM architecture. Stage 1: Dense captioning via curriculum-trained VLM. Stage 2: Multimodal causal head to predict causal adjacency matrix with joint training and class imbalance handling.

stratified by event count, blind to all LLM labels (see Appendix G). Our released labels achieve edge precision 0.92 and recall 0.82. Agreement on non-edges is $\kappa = 0.79$, and the graph-level inter-annotator agreement is 0.83.

Our experiments provide direct empirical evidence: Gemini 2.5 Pro and GPT-5 fail in predicting causal relationships in long-form videos (Table 2), even when given structured prompts and video. By contrast, LLM-first annotation (from ground-truth captions) applies hard structural constraints (temporal ordering, cardinality), and is then validated visually via EMScore + VLM+LLM verification + human judgment. This follows and goes beyond well-established pipelines of NarrativeBridge [3] and MECD [8], both using LLM-generated annotations. We also independently re-score all the human-verified samples (from step-3) using DeepSeek-R1 [51] as the judge in place of Llama-3.3-70B from step 2, obtaining near-identical mean scores and high cross-LLM agreement (Pearson $r = 0.90$), confirming the validation scores are not an artefact of within-family agreement but within 2 LLMs and 3 humans.

3.3 Causal-VLM Architecture

Our architecture consists of two training stages shown in Figure 3. Stage 1 trains a dense video captioning model using curriculum learning with DINOv2-based frame sampling to extract semantically coherent video segments. Stage 2 adds a multimodal causal pipeline predicting event relationships. This staged approach outperforms joint training because causal prediction requires high-quality event representations established before causal learning begins.

DINOv2 Frame Sampling. Video preprocessing requires semantically-aware frame selection beyond uniform temporal or random sampling. We employ pretrained DINOv2 [52] ViT-S/14 for scene detection, extracting features from frames and then, scene boundaries are identified where cosine similarity between consecutive frame features falls below threshold $\theta = 0.8$ (see Appendix D.1):

$$\text{boundary at } t \text{ if } \frac{\mathbf{f}_t \cdot \mathbf{f}_{t+1}}{\|\mathbf{f}_t\|\|\mathbf{f}_{t+1}\|} < \theta \tag{1}$$

where $\mathbf{f}_t$ denotes the feature vector for frame t. Detected scenes are balanced to 64 frames: if too few scenes exist, we split the longest scenes at minimum-similarity frames; if too many, we merge the most similar adjacent scenes.

3.3.1 Stage 1: Dense Video Captioning with Curriculum Training:

Our architecture combines a vision Perception encoder-L/14 [53], a vision-language projection layer, and Llama-3.2-3B-Instruct [18] language model. We employ three-phase curriculum training [23] that progressively unfreezes model capacity to improve convergence, stability and generalization.

Phase 1 trains only the projection layer while freezing vision encoder and language model. This learns the mapping between pretrained visual features and language model input space without risking forgetting either components pretrained knowledge. The projector applies two linear layers separated

by a GELU activation to each visual patch token, followed by adaptive spatial average pooling. Its weights are randomly initialized and learned during Stage 1. We optimize the autoregressive language modeling objective:

$$\mathcal{L}_{\text{caption}} = -\frac{1}{T}\sum_{t=1}^{T} \log p(w_t | w_{<t}, \mathbf{v}) \tag{2}$$

where w_t is the t-th token, $w_{<t}$ represents preceding tokens, and $\mathbf{v}$ denotes projected visual features.
Phase 2 unfreezes the vision encoder while continuing projection layer training, enabling domain adaptation to ActivityNet and YouCook2 videos.
Phase 3 enables full end-to-end fine-tuning by unfreezing all model components, including the language model. This allows the entire architecture to adapt jointly to the dense video captioning task. The resulting model generates good-quality dense captions.

3.3.2 Stage 2: Multimodal Causal Learning:

Architectural Motivation. We initially experiment with a pure visual baseline without textual conditioning but achieve only 0.34 F1 score (see Section 4.4). The core issue is that visual features alone cannot capture semantic causal prerequisites and require language reasoning.
We therefore design a multimodal architecture fusing visual features with language semantic information. As shown in Stage 2 of Figure 3, the architecture takes two inputs: frame-level visual features $\mathbf{V} \in \mathbb{R}^{B\times F\times D_v}$ from Stage 1 encoder, and similarly, language embeddings $\mathbf{T} \in \mathbb{R}^{B\times N_{\max}\times D_t}$ of event captions, obtained by passing each caption through the Stage-1 language model and mean-pooling its last-layer hidden states into one vector per event. Here B is the batch size, F is sampled frames using DINOv2 [52], $N_{\max} = 20$ is the maximum number of events per video, D_v is the visual feature dimension, and D_t is the text embedding dimension. It outputs pairwise causal edges; parallel patterns (A,B→C) emerge as correlated edge predictions across the shared contextual representations from the transformer encoder.
Event-Level Pooling. For each video with N events, we aggregate the DINOv2-sampled frame features to event-level representations using the event boundaries generated by Stage 1. Let $[s_i, e_i]$ be the predicted boundaries of event i, τ_t the timestamp of sampled frame t, and $\mathcal{S}_i = \{t : s_i \le \tau_t \le e_i\}$. We average the frame features inside each predicted segment:

$$\mathbf{V}_{\text{event},i} = \frac{1}{|\mathcal{S}_i|}\sum_{t\in\mathcal{S}_i} \mathbf{V}_t, \qquad i = 1, \ldots, N \tag{3}$$

This gives $\mathbf{V}_{\text{event}} \in \mathbb{R}^{N\times D_v}$, one representation per event, and overlapping events share frames.
Cross-Modal Fusion. We project both modalities to a common hidden space of dimension H:

$$\mathbf{V}' = \text{GELU}(\text{LayerNorm}(\mathbf{W}_v \mathbf{V}_{\text{event}})) \tag{4}$$

$$\mathbf{T}' = \text{GELU}(\text{LayerNorm}(\mathbf{W}_t \mathbf{T}[:N])) \tag{5}$$

where $\mathbf{W}_v \in \mathbb{R}^{H\times D_v}$ and $\mathbf{W}_t \in \mathbb{R}^{H\times D_t}$ are learned projection matrices. The projected features fuse through multi-head cross-attention where visual features query language representations:

$$\mathbf{F}, _ = \text{MultiHeadAttention}(\mathbf{V}', \mathbf{T}', \mathbf{T}') \tag{6}$$

This allows each event's visual representation to selectively retrieve relevant semantic information from its textual description. We add a residual connection preserving the original visual signal, resulting in $\mathbf{E}$:

$$\mathbf{E} = \mathbf{V}' + \mathbf{F} \tag{7}$$

Transformer Encoder. The fused representations are contextualized through a two-layer transformer encoder:

$$\mathbf{C} = \text{TransformerEncoder}(\mathbf{E}) \tag{8}$$

Each layer consists of a multi-head self-attention followed by a position-wise feedforward network (hidden dimension $2H$), with layer normalization and residual connections. This contextual encoding is necessary because an event's causal role depends on surrounding events. For example, "season chicken" is only causal if followed by "coat chicken"; in isolation, it might be preparatory or independent.
Pairwise MLP. For each ordered pair (i, j), we construct pairwise features by concatenating contextual representations:

$$\mathbf{P}_{ij} = [\mathbf{C}_i; \mathbf{C}_j] \in \mathbb{R}^{2H} \tag{9}$$

These pairwise representations capture both events' characteristics and implicitly their temporal relationship (order matters: $\mathbf{P}_{ij} \neq \mathbf{P}_{ji}$). We process through a three-layer MLP with decreasing dimensions:

$$\mathbf{h}^{(1)} = \text{GELU}(\text{LayerNorm}(\mathbf{W}^{(1)}\mathbf{P}_{ij})) \tag{10}$$

$$\mathbf{h}^{(2)} = \text{GELU}(\mathbf{W}^{(2)}\mathbf{h}^{(1)}) \tag{11}$$

$$z_{ij} = \mathbf{w}^{(3)^T}\mathbf{h}^{(2)} \tag{12}$$

where $\mathbf{W}^{(1)} \in \mathbb{R}^{H\times 2H}$, $\mathbf{W}^{(2)} \in \mathbb{R}^{H/2\times H}$, and $\mathbf{w}^{(3)} \in \mathbb{R}^{H/2}$. Final prediction applies sigmoid activation and enforces temporal causality through masking:

$$A_{ij} = \begin{cases} \sigma(z_{ij}) & \text{if } i < j \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the causal adjacency matrix with A_{ij} indicating the probability that event i causes event j. The mask restricts predictions to forward, acyclic graphs, matching the label space.

3.3.3 Training with Extreme Class Imbalance

Class imbalance varies substantially across videos. A video with 10 events and 2 causal pairs has 43 negatives vs. 2 positives (21:1); at the tail (21 events, 1 causal pair), ratios exceed 200:1. This causes standard BCE to collapse for sparse videos, achieving high accuracy by predicting "no causality" everywhere. We address this with three techniques: **(i) Uncapped positive weighting** $w_{\text{pos}} = n_{\text{neg}}/n_{\text{pos}}$ (up to 50×); **(ii) Hard negative mining** [25] retaining only the top-20% most-confident negatives per video (selected via grid search 5%–40%); **(iii) Asymmetric false negative penalty** $\lambda_{\text{fn}} = 10$ to bias toward recall, since missing a true causal dependency is more harmful than a false positive. Uncapped weighting corrects the prior imbalance; HNM stabilises training by reducing the effective negative count to 20% of n_{neg}; and the FN penalty counteracts the residual tendency to suppress rare positives. Without all three, standard BCE collapses to 0.31 F1 (Table 4c, Std), and removing any single component degrades F1 by 0.14–0.19. Full equations are in Appendix A. The final loss and joint objective are:

$$\mathcal{L}_{\text{causal}} = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega} w_{ij}\cdot M_{ij}\cdot \text{BCE}(A_{ij}, Y_{ij}), \quad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{caption}} + \lambda\cdot\mathcal{L}_{\text{causal}} \tag{14}$$

where $\lambda = 2$ and $\Omega = \{(i,j) : i < j\}$. Joint training improves both causal prediction and caption quality (see Section 4.4).

4 Experiments

4.1 Setup

Datasets. We evaluate on four datasets. For dense causal captioning, we use ActivityNet Captions [16] and YouCook2 [17] with our generated causal annotations (DCC benchmark; full statistics in Appendix B). We follow the standard ActivityNet Captions and YouCook2 splits. We also report results on MSR-VTT CTN [3], MSVD-CTN [3] and NExT-QA [4].

Implementation. We implement all models in PyTorch [54] with mixed-precision training (bfloat16) on NVIDIA H200 GPUs using a batch size of 16. We preprocess videos using DINOv2 scene detection, sampling 64 frames per video. Main results are means over 3 seeds, with standard deviation $\pm$0.010 F1 and $\pm$0.8 CIDEr.

Caption Quality Metrics. We evaluate dense captioning using CIDEr [55], ROUGE-L [56], and SPICE [57], computing metrics per-event. CIDEr is computed per event against IoU-matched ground-truth segments using predicted proposals.

Causal Prediction Metrics. We treat predicted relationships as directed graphs and compute precision, recall, and F1 against ground truth. A predicted link (i, j) is correct if $A_{ij} > \tau^*$ and the ground truth contains the same link. $\tau^* = 0.5$ is selected by grid search. Predicted events are matched to ground-truth events by temporal IoU ≥ 0.5.

Inference Pipeline. Stage 1 generates event descriptions $\{s_i\}_{i=1}^N$ with timestamps $\{t_i\}_{i=1}^N$, which are encoded to produce text embeddings $\mathbf{T}$. Stage 2 fuses these with visual features $\mathbf{V}$ to predict the causal adjacency matrix $\mathbf{A}$.

4.2 Main Results

Table 2 shows results on ActivityNet [16] and YouCook2 [17]. Our approach achieves 0.73 F1 on causal prediction, outperforming visual-only baselines (0.34 F1) and fine-tuned VLMs (0.27-0.33 F1).

Causal supervision improves caption quality: +6.3 CIDEr on ActivityNet, +7.3 on YouCook2. Joint training also improves temporal localization (+0.8 at IoU 0.5), indicating that causal supervision helps segment event boundaries. Fine-tuned VLMs learn output formats but not the causality. Closed-source models (Gemini 2.5 Pro, GPT-5) show low performance on causal relationship prediction. These zero-shot rows motivate the task and are not a like-for-like comparison with a supervised model. A format-tolerant parser that accepts free-text edge lists and repaired JSON changes their F1 by at most 0.01, so the low scores are not a parsing artefact. The low CIDEr reflects caption verbosity mismatch with terse ground-truth annotations. Five-shot prompting improves these models by 0.05–0.1 F1 and 5–8 CIDEr points but remains below fine-tuned baselines; we report zero-shot for fair comparison. We also report AUPRC 0.687/0.616 on ActivityNet/YouCook2 and per-video macro-F1 is within 0.04 of micro-F1 for our model (see Appendix I). CausalVLM transfers to VideoQA settings: **+6.4** on NExT-QA causal-Q accuracy and 0.82 oracle F1 vs MECD's 0.71 [8] (see Appendix F and E respectively).

Caption-to-LLM baseline. Passing Stage-1 captions and timestamps to an instruction-tuned LLM with the annotation prompt, which then predicts only the graph, reaches at best 0.487/0.369 F1 (Llama-3.3-70B), so Causal-VLM leads by 0.24/0.27 F1. Our 11.6M-parameter head is also four orders of magnitude faster. With ground-truth captions, an oracle setting unavailable at inference, the gap narrows (0.822/0.730 vs 0.79/0.74), but the LLM loses 0.34/0.36 F1 on predicted captions and timestamps while our head loses 0.06/0.10 on ActivityNet/YouCook2 (Appendix H).

Table 2: Dense causal captioning results. F1: causal prediction, C: CIDEr, R-L: ROUGE-L, S: SPICE, R@X: Recall at IoU threshold X. Three groups: (i) open-source models (ZS=zero-shot, FT=fine-tuned on DCC); (ii) closed-source ZS only; (iii) proposed. The LLM row predicts only the graph from Stage-1 captions. C-VLM: mean of 3 seeds ($\pm$0.010 F1, $\pm$0.8 CIDEr).

	ActivityNet						YouCook2					
Method	**F1**	**C**	**R-L**	**S**	**R@.5**	**R@.7**	**F1**	**C**	**R-L**	**S**	**R@.5**	**R@.7**
Open-source, task-adapted (ZS = zero-shot; FT = fine-tuned on DCC)												
PLM-3B [12] (ZS)	0.18	44.1	29.3	15.8	73.8	48.5	0.18	40.2	27.1	14.6	76.0	52.1
PLM-3B [12] (FT)	0.29	45.7	30.1	16.4	74.2	48.9	0.30	41.8	27.8	15.2	76.5	52.6
InternVL3-4B [13] (ZS)	0.23	45.3	29.8	16.1	74.0	48.7	0.23	41.4	27.5	15.0	76.2	52.3
InternVL3-4B [13] (FT)	0.31	46.8	30.5	16.7	74.5	49.1	0.33	42.9	28.3	15.6	76.8	52.8
InternVL3-8B [13] (FT)	0.30	47.2	30.1	15.4	74.3	49.1	0.32	42.5	27.5	15.3	76.9	52.9
Qwen3-VL-8B [20] (FT)	0.27	48.4	29.5	16.3	74.5	49.4	0.28	43.2	27.9	15.7	77.1	53.2
Closed-source, zero-shot (ZS) only												
Gemini 2.5 Pro [21] (ZS)	0.16	12.6	15.9	10.5	24.8	12.1	0.10	16.0	17.4	14.6	30.6	11.6
GPT-5 [22] (ZS)	0.12	13.9	16.7	11.6	27.4	13.8	0.11	17.2	17.9	14.8	33.1	12.1
Stage-1 captions to LLM (graph only)												
Llama-3.3-70B [18]	0.49	–	–	–	–	–	0.37	–	–	–	–	–
Proposed architecture												
Visual-only	0.34	42.3	28.6	15.2	72.1	47.8	0.36	38.7	26.4	14.1	75.3	51.2
C-VLM Stage-1 (Ours)	–	45.7	30.8	17.2	73.8	49.7	–	38.8	28.0	15.8	77.5	53.9
C-VLM Full (Ours)	**0.73**	**52.0**	**33.2**	**18.9**	**74.6**	**50.3**	**0.64**	**46.1**	**30.1**	**17.3**	**78.3**	**54.7**

Comparison on CTN Datasets. Table 3 compares caption quality on CTN datasets [3]. Our approach outperforms the NarrativeBridge by 3-4 CIDEr points on both MSR-VTT CTN and MSVD CTN. NarrativeBridge uses separate cause and effect encoders for causal-temporal captioning, while our model uses a single encoder. Training on longer videos with complex causal structures (ActivityNet, YouCook2) provides representations that transfer to shorter videos with single causal relationships.

Table 3: Caption quality on MSR-VTT CTN and MSVD CTN. C: CIDEr, R-L: ROUGE-L, S: SPICE.

	MSR-VTT CTN			MSVD CTN		
Method	**C**	**R-L**	**S**	**C**	**R-L**	**S**
GIT [58]	32.4	24.5	13.7	45.6	27.5	15.6
Video-LLaVA [40] (FT)	20.0	22.2	13.1	36.1	25.6	16.1
ShareGPT4Video [59] (FT)	20.8	23.0	13.3	31.7	25.3	14.9
NarrativeBridge [3]	49.9	27.9	15.8	63.5	31.5	19.3
Causal-VLM (Ours)	**53.2**	**29.4**	**16.8**	**67.8**	**32.9**	**20.1**

4.3 Qualitative Results

Figure 4 shows predictions across three causal structures. In the dog-frisbee example (a), E3 (camera falling) depends on *both* E1 (person walking with dog) and E2 (frisbee thrown), a joint dependency where both incoming edges must be correct. PLM-3B produces invalid output for E1 and misclassifies E2 as a pure effect, missing the dual role entirely; InternVL-4B correctly identifies E1 but fails to



Figure 4: Qualitative comparison with PLM-3B [12] and InternVL-4B [13]. Green checkmarks: correct; red crosses: errors. (a) Camera falling (E3) requires both E1 and E2. (b) Linear causal relationship. (c) Parallel chains: sauce (E1–E3) and potato (E4–E5) both required for plating (E6).

recognise E2's causal role. In the skiing example (b), a linear chain requires each event to act as both cause and effect. Both baselines misclassify the middle event E2 as a pure effect, losing the chain structure; our model identifies all three links correctly. In the cooking example (c), two independent preparation chains (E1–E3: sauce; E4–E5: potatoes) converge at the final plating step E6. PLM-3B achieves 1/6 correct predictions; InternVL-4B achieves 3/6 but misses the parallel structure. Across all three cases, baseline failures reduce to two patterns: (1) over-predicting *effect* labels for events that also initiate subsequent steps, and (2) producing invalid outputs for intermediate events where neither a pure cause nor a pure effect label fits. Additional examples across eight videos are in Appendix K.

Table 4: Ablation studies on ActivityNet (Section 4.4). **(a)** Architecture: each column removes one component, text emb. (T), cross-modal fusion (F), transformer encoder (E), temporal mask (M). **(b)** Curriculum: joint training from scratch (Jt) vs phases 1+2 only (P1+2). **(c)** Training loss: removing hard neg. mining (H), uncapped weighting (U), FN penalty (P), or standard BCE (Std). **(d)** Frame sampling: DINOv2 vs random (Rnd) and periodic (Per). **(e)** F1 by event type. **(f)** Oracle: GT timestamps and captions given to Stage 2. **(g)** Modality: same arch/loss, input varies. **(h)** Grounding: shuffled visual features, permuted or corrupted captions, uniform pooling. In (f), T/C: timestamps/captions, GT or predicted (Pr).

(a) Architecture

M.	Full	T	F	E	M
F1	**.73**	.34	.47	.52	.58
C	**52**	42	50	50	51
R@.5	**74.6**	72	73	74	74

(b) Curriculum

M.	Full	Jt	P1+2
F1	**.73**	.68	.70
C	**52**	49.2	50.3
R@.5	**74.6**	73.9	74.2

(c) Training

M.	Full	H	U	P	Std
F1	**.73**	.59	.54	.57	.31
C	**52**	50	50	50	48
R@.5	**74.6**	74	73	74	73

(d) Frame Samp.

M.	DINO	Rnd	Per
F1	**.73**	.67	.69
C	**52**	45.8	48.2
R@.5	**74.6**	71.3	72.8

(e) Event Type

Type	P	R	F1
Cause	.82	.75	.78
Effect	.76	.71	.73
Causal	.71	.63	.67
Indep.	.84	.79	.81

(f) Oracle

T	C	AN	YC
GT	GT	.79	.74
GT	Pr	.74	–
Pr	GT	.77	–
Pr	Pr	.73	.64

(g) Modality

Input	AN	YC
Text-only	.31	.24
Visual-only	.34	.36
Multimodal	**.73**	**.64**

(h) Grounding

Input (AN)	F1
Shuf. V	.52
Perm. C	.55
Corr. 25/50%	.69 / .55
Unif. pooling	.67

4.4 Ablation Studies

Table 4 ablates each design dimension on ActivityNet.

(a) Architecture. Removing text embeddings causes the drop (0.73→0.34 F1), confirming that visual features alone cannot resolve semantically similar events (e.g. "season chicken" vs "season flour"). Cross-modal fusion contributes 0.26 F1 by aligning visual and semantic spaces, the transformer encoder adds 0.21 F1 by contextualising each event relative to its neighbours, and temporal masking adds 0.15 F1 by enforcing causal directionality.

(b) Curriculum. Training jointly from the start yields 0.68 F1, compared to 0.73 for our three-phase curriculum. Stopping after Phase 2 achieves 0.70 F1. The gap shows that Phase 3 end-to-end fine-tuning is necessary for the captioning and causal components to co-adapt; causal supervision cannot improve caption quality unless the VLM backbone is also allowed to update.

(c) Training. Standard BCE collapses to 0.31 F1 due to 9:1–75:1 class imbalance. Each imbalance-handling technique is necessary: removing hard negative mining (H) drops to 0.59, removing

uncapped weighting (U) drops to 0.54, and removing the FN penalty (P) drops to 0.57. Together they recover the full 0.73 F1.

(d) Frame sampling. DINOv2 scene-aware sampling outperforms random (+0.06 F1, +6.2 CIDEr, +3.3 R@0.5) and periodic (+0.04 F1, +3.8 CIDEr, +1.8 R@0.5) alternatives. Semantically coherent frame boundaries improve event localisation, which in turn benefits causal prediction, consistent with the localization-causality correlation observed in Section 4.4.

(e) Event types. Independent events (no causal role) are easiest at 0.81 F1, followed by cause-only (0.78) and effect-only (0.73) events. "Causal" events, which are simultaneously caused by prior events and enable subsequent ones, are hardest at 0.67 F1, requiring the model to reason about both incoming and outgoing dependencies.

(f) Oracle. Replacing Stage 1 predictions with ground-truth timestamps and captions yields 0.79 F1 on ActivityNet and 0.74 on YouCook2, versus 0.73 and 0.64 for the full pipeline. The 6–10 point gap quantifies Stage 1 error propagation; the larger gap on YouCook2 reflects its denser causal chains, where localisation errors compound more severely. With only captions or only timestamps predicted, F1 is 0.74 or 0.77, so caption errors dominate the gap.

(g) Modality. Using the same architecture and loss but zeroing visual features gives 0.31 F1 (text-only) and zeroing text gives 0.34 F1 (visual-only). The full multimodal model achieves 0.73 F1, with the +0.42 gap over text-only confirming visual grounding is necessary.

(h) Grounding. Without retraining, shuffling visual features across videos lowers F1 to 0.52 and permuting captions to 0.55, far beyond the ± 0.010 seed deviation, so the head uses the video. Uniform pooling gives 0.67.

5 Limitations

While Causal-VLM advances dense causal captioning, some limitations remain. First, DCC labels are generated from an LLM reading human-annotated captions and the LLM can inherit its procedural priors and omissions in the captions. The human blind audit estimates label quality on 200 of the 21.6K videos, and defensible edges missing from the labels count as false positives. As specified in Section 3.2, edges capture observed prerequisite structure: necessity is not verified counterfactually, goals and intentions are not represented, and cyclic relations are excluded. Second, ActivityNet and YouCook2 are mostly English-language web videos and YouCook2 covers only cooking, so DCC inherits their domain bias. Third, model performance degrades on causal chains of five or more hops and across domains: joint training on both datasets gives 0.60 F1 on YouCook2, against 0.64 in-domain and 0.47 when trained on ActivityNet only, so part of the causal structure is domain-specific. Fourth, Stage 1 localisation errors propagate to Stage 2: the oracle gap (0.79 vs 0.73 F1 on ActivityNet) shows that further gains in captioning quality will directly benefit causal prediction. Finally, causal F1 depends on IoU event matching, and per-event CIDEr is not comparable with paragraph-level results.

6 Conclusion

We introduce DCC, the first large-scale benchmark for dense causal captioning, comprising 21.6K videos and 85K events across ActivityNet and YouCook2, validated through multimodal alignment, VLM verification, and human judgment, including a blind full-graph audit. Building on DCC, Causal-VLM demonstrates that explicitly supervising causal relationships, through a two-stage architecture combining curriculum-trained dense captioning with a multimodal causal head, allows models to predict causal relationships between events in long-form videos. The result is a system that not only predicts causal relationships effectively (0.73/0.64 F1) but simultaneously improves caption quality (+6.3/+7.3 CIDEr) and temporal localisation of the events, showing that causal learning is a broadly beneficial training signal rather than a narrow auxiliary task. Six VLM baselines and a caption-to-LLM pipeline fall well short without specialised supervision. We expect DCC and Causal-VLM to enable downstream applications in instructional video systems (prerequisite identification), robotics (task planning and failure diagnosis), and accessibility tools (explanatory audio descriptions for visually impaired users). Future work should explore extensions to counterfactual and interventional reasoning, testing the necessity of these edges, and application to longer-horizon causal chains in domains beyond cooking and activity videos.

Acknowledgments and Disclosure of Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) and BBC Research & Development through the Prosperity Partnership grant EP/V038087/1, "BBC

Prosperity Partnership: Future Personalised Object-Based Media Experiences Delivered at Scale Anywhere", and by the UK Arts and Humanities Research Council (AHRC) through the CoSTAR National Lab (AH/Y001060/1). We thank the annotators who took part in the human validation study and the blind full-graph audit.

References

[1] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[2] Jeroen Berrevoets, Krzysztof Kacprzyk, Zhaozhi Qian, and Mihaela van der Schaar. Causal deep learning. *arXiv preprint arXiv:2303.02186*, 2023.

[3] Asmar Nadeem, Faegheh Sardari, Robert Dawes, Syed Sameed Husain, Adrian Hilton, and Armin Mustafa. Narrativebridge: Enhancing video captioning with causal-temporal narrative. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=bBoetBIN2R`.

[4] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021.

[5] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[6] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025.

[7] Xuchen Li, Xuzhao Li, Shiyu Hu, Kaiqi Huang, and Wentao Zhang. Causalstep: A benchmark for explicit stepwise causal reasoning in videos. *arXiv preprint arXiv:2507.16878*, 2025.

[8] Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, et al. Mecd: Unlocking multi-event causal discovery in video reasoning. *Advances in Neural Information Processing Systems*, 37:92554–92580, 2024.

[9] Tieyuan Chen, Huabin Liu, Yi Wang, Yihang Chen, Tianyao He, Chaofan Gan, Huanyu He, and Weiyao Lin. Mecd+: Unlocking event-level causal graph discovery for video reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[12] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025.

[13] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023.

[15] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, 2024.

[16] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[17] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[18] The Llama 3 Herd of Models. Technical Report, Meta, July 2024. URL `https://ai.meta.com/research/publications/the-llama-3-herd-of-models/`.

[19] Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17929–17938, 2022.

[20] Qwen Team. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL `https://arxiv.org/abs/2511.21631`.

[21] 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. URL `https://arxiv.org/abs/2507.06261`.

[22] OpenAI. Gpt-5 system card, 2025. URL `https://cdn.openai.com/gpt-5-system-card.pdf`. Published August 13, 2025.

[23] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning (ICML)*, pages 41–48, 2009.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2980–2988, 2017.

[25] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 761–769, 2016.

[26] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21273–21282, 2022.

[27] Jiayuan Mao, Xuelin Yang, Xikun Zhang, Noah Goodman, and Jiajun Wu. CLEVRER-Humans: Describing physical and causal events the human way. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

[28] Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. Black swan: Abductive and defeasible video reasoning in unpredictable events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24201–24210, 2025.

[29] Paritosh Parmar, Eric Peh, Ruirui Chen, Ting En Lam, Yuhan Chen, Elston Tan, and Basura Fernando. CausalChaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

[30] Miaoyu Li, Qin Chao, and Boyang Li. Two causally related needles in a video haystack. In *Advances in Neural Information Processing Systems*, 2025.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[32] 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 *International Conference on Machine Learning (ICML)*, 2021.

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[34] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[35] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[36] Aishwarya Kamath, Mannat Singh, Yann Le, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[37] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022.

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. *URL https://arxiv. org/abs/2310.03744*, 3(4):5, 2023.

[39] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[40] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024.

[41] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1207–1216, 2019.

[42] Chih-Yao Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 73–90. Springer, 2020.

[43] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[44] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[45] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[48] Judea Pearl. Causality: Models, reasoning, and inference. In *Cambridge University Press*, 2009.

[49] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[50] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

[51] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[52] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[53] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[55] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[56] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[57] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.

[58] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[59] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024.

[60] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

[61] Wei Dai, Alan Luo, Zane Durante, Debadutta Dash, Arnold Milstein, Kevin Schulman, Ehsan Adeli, and Li Fei-Fei. Towards fine-grained video question answering. *arXiv preprint arXiv:2503.06820*, 2025.

[62] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *NeurIPS*, 2023.

[63] Juhong Min, ..., and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13235–13245, 2024.

[64] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15405–15416, 2023.

A Implementation Details

This appendix section supplements Section 3 (*Method*) of the main paper. Tables 5–7 provide the full hyperparameter configurations required for reproducibility. Section A.3 gives the complete training-loss equations.

A.1 Stage 1 Training Configuration

Stage 1 trains the dense video captioning component described in §3.3 of the paper using a three-phase curriculum learning. Each phase progressively unfreezes model capacity: Phase 1 trains only the projection layer to align visual and language embedding spaces without disturbing pretrained knowledge; Phase 2 unlocks the vision encoder for domain adaptation to ActivityNet and YouCook2; Phase 3 enables full end-to-end fine-tuning with a language model LR two orders of magnitude smaller than the projector to prevent catastrophic forgetting.

Table 5: Stage 1 curriculum training configuration (Section 3.3, Phase 1–3 of main paper). "frozen" indicates the component's parameters are not updated in that phase. The substantially lower LLM LR in Phase 3 (1e-6 vs 1e-4 for the projector) preserves pretrained language knowledge while allowing task-specific adaptation.

Parameter	Phase 1	Phase 2	Phase 3
Epochs	5	10	10
Batch size	16	16	16
Projection LR	1e-4	1e-4	5e-5
Vision encoder LR	frozen	1e-5	1e-5
Language model LR	frozen	frozen	1e-6
Weight decay (LLM)	–	–	0.01
Gradient clip norm	1.0	1.0	1.0
Warmup steps	500	500	1000
LR schedule	cosine	cosine	cosine
Optimizer	AdamW	AdamW	AdamW
Precision	bfloat16	bfloat16	bfloat16

We use AdamW [60] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and gradient clipping at max norm 1.0 throughout all phases. These settings follow standard practice for large VLM fine-tuning.

A.2 Stage 2 Training Configuration

Stage 2 adds the multimodal causal head (Section 3.3 of the paper, Stage 2) and trains jointly with the captioning model. The two-phase schedule is critical: the first 10 epochs freeze the captioning model so the causal head can learn basic dependency patterns without interfering with captioning representations; the final 10 epochs enable end-to-end refinement. The loss balance weight $\lambda = 2.0$ (main paper Eq. 14) was selected by grid search over $\{0.5, 1.0, 2.0, 5.0, 10.0\}$ on the ActivityNet validation set, maximising causal F1 without degrading caption CIDEr below the Stage 1 baseline. The hard negative mining ratio of 20% was chosen from $\{5\%, 10\%, 20\%, 30\%, 40\%\}$ (see Appendix D.3 for the full sweep).

Table 6: Stage 2 causal head training configuration (Section 3.3, Stage 2 of main paper). During the frozen phase, only the causal head and loss parameters are updated; in the end-to-end phase, all components are jointly fine-tuned with carefully scaled learning rates to avoid catastrophic forgetting of captioning ability.

Parameter	Frozen (Ep. 1–10)	End-to-End (Ep. 11–20)
Causal head LR	1e-4	1e-4
Vision projector LR	frozen	1e-5
Vision encoder LR	frozen	1e-6
Language model LR	frozen	1e-6
Loss balance λ	2.0	2.0
Pos. weight (uncapped)	$n_{\text{neg}}/n_{\text{pos}}$	$n_{\text{neg}}/n_{\text{pos}}$
Hard neg. ratio	20%	20%
False neg. penalty λ_{fn}	10	10
Batch size	16	16
Gradient accumulation	4	4

A.3 Full Training Loss Equations

The following three equations define the class-imbalance handling techniques introduced in Section 3.3 (*Training with Extreme Class Imbalance*) of the main paper. The main paper presents only the combined causal loss and joint objective (Eq. 14); the component-level equations are provided here for reproducibility.

(i) Uncapped Positive Weighting (main paper Section 3.3, "Uncapped Positive Weighting"). Unlike standard practice that clamps positive weights at 5–10×, we use the true per-video imbalance ratio without a ceiling. This can exceed $50\times$ in sparse videos; training stability is maintained by the hard negative mining that reduces the effective number of negatives:

$$w_{\text{pos}} = \frac{n_{\text{neg}}}{n_{\text{pos}}} \tag{15}$$

(ii) Hard Negative Mining Mask (main paper Section 3.3, "Hard Negative Mining"). For each video we retain all positive pairs and only the top-20% most-confusing negatives, those where the model currently assigns the highest causal probability, discarding easy negatives that provide no gradient signal:

$$M_{ij} = \begin{cases} 1 & \text{if } Y_{ij} = 1 \\ 1 & \text{if } Y_{ij} = 0 \text{ and } A_{ij} \in \text{top-}20\% \text{ of negatives} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

(iii) Asymmetric False Negative Penalty (main paper Section 3.3, "Asymmetric False Negative Penalty"). Missing a true causal link is costlier than predicting a spurious one, particularly in instructional videos where a missed prerequisite could mislead a viewer or a robotic planner. We apply penalty factor $\lambda_{\text{fn}} = 10$ (see Appendix D.4 for the sweep) to false negative positions:

$$w_{ij} = \begin{cases} \lambda_{\text{fn}} \cdot w_{\text{pos}} & \text{if } Y_{ij} = 1 \text{ and } A_{ij} < 0.5 \\ w_{\text{pos}} & \text{if } Y_{ij} = 1 \text{ and } A_{ij} \geq 0.5 \\ 1 & \text{if } Y_{ij} = 0 \end{cases} \tag{17}$$

A.4 Causal Head Architecture Dimensions

Table 7 provides the exact layer dimensions for the multimodal causal head described in Section 3.3 (Stage 2) of the main paper. The hidden dimension $H = 512$ is shared across cross-modal fusion and the transformer encoder. The pairwise MLP processes each ordered event pair (i, j) by compressing a $2H$-dimensional concatenation to a scalar causal probability via three decreasing-dimension layers.

Table 7: Causal head architectural dimensions (Stage 2 of main paper, Section 3.3). The pairwise MLP input is $[\mathbf{C}_i; \mathbf{C}_j] \in \mathbb{R}^{2H}$ (Eq. 9 of main paper); it progressively compresses to a scalar logit.

Component	Dimension
Cross-modal Fusion (Eq. 6–7 of main paper)	
Attention heads	8
Head dimension ($H/8$)	64
Dropout	0.1
Transformer Encoder (Eq. 8 of main paper)	
Layers	2
Attention heads	8
FFN hidden dim ($2H$)	1024
Dropout	0.1
Pairwise MLP (Eqs. 9–13 of main paper)	
Layer 1 output (H)	512
Layer 2 output ($H/2$)	256
Layer 3 output (scalar)	1
Dropout (Layer 1)	0.2

B DCC Benchmark Dataset Statistics

This section supplements Section 3.2 (*Dense Causal Captions (DCC) Benchmark Dataset*) of the main paper. Table 8 provides complete benchmark statistics including train/val splits, event-type distributions, and causal chain length profiles. The chain-length distribution is particularly important

for understanding why YouCook2 is harder than ActivityNet: its 75.2% four-hop-or-more chains create densely interconnected causal graphs that require predicting across many event boundaries.

Table 8: DCC benchmark statistics (Section 3.2 of main paper). The contrasting distributions between ActivityNet (sparse, short chains; diverse activities) and YouCook2 (dense, long chains; procedural cooking) explain the performance gap between datasets observed in Table 2 of the main paper (0.73 vs 0.64 F1 for Causal-VLM).

Statistic	**ActivityNet**	**YouCook2**
Overall		
Total videos	19,796	1,790
Total events	71,532	13,812
Total causal pairs	57,813	16,375
Train / Test / Val split	10,009 / 4870 / 4,917	1,333 / 457
Mean duration (seconds)	118	321
Events per Video		
Mean	3.61	7.72
Max	21	16
Causal Pairs per Video		
Mean	2.92	9.15
Max	45	91
Event Type Distribution		
Cause (initiates others)	27.5%	23.0%
Effect (results from others)	33.3%	15.2%
Causal (Cause and Effect mediator)	36.4%	61.6%
Independent (no links)	2.9%	0.2%
Causal Chain Length		
Single-hop (A→B)	16.6%	3.3%
Two-hop (A→B→C)	50.4%	7.8%
Three-hop	18.1%	13.7%
Four-hop or more	14.9%	75.2%

YouCook2 is dominated by *Causal* events (61.6%) because cooking steps both depend on prior steps *and* enable subsequent ones. Its extreme long-chain prevalence (75.2% four-hop+) reflects multi-step ingredient transformations: chop → sauté → season → simmer → reduce → plate.

C DCC Prompts

This section supplements Section 3.2(*DCC Benchmark Dataset Generation*) of the main paper, providing the complete prompts used for LLM annotation (Section C.1) and VLM+LLM validation (Section C.2). These prompts implement the five structured constraints described in Section 3.2: explicit event-type definitions, temporal consistency, linguistic cue guidance, cardinality constraints, and JSON output format.

C.1 LLM Annotation Generation Prompt

The following prompt drives Llama-3.3-70B [18] to generate the causal adjacency annotations described in Section 3.2. Hard constraints in the prompt (items 3–5) prevent the common failure mode of generating more events than the input captions describe. Guideline 1 defines a cause as an event that "directly leads to or influences" other events. Here this means an observed prerequisite dependency, as described in Section 3.2 of the main paper; it does not assert interventional necessity.

```
You are an expert in analyzing video narratives
and identifying cause-effect relationships in sequences
of events. Convert the following video caption data into
a structured narrative format that shows cause-effect
relationships between events.
Input video caption data:
{input_json}
Generate a single JSON object as output with exactly
this structure:
{
    "duration": float,
```

```
    "events": [
        {
            "event_id": int,
            "timestamp": [start, end],
            "description": "text",
            "chain": {
                "is_cause": boolean,
                "is_effect": boolean,
                "cause_event": [event_ids],
                "effect_event": [event_ids]
            },
            "event_type": "cause,effect,causal,independent"
        }
    ]
}

Guidelines for analysis:
0. Count total number of events in input first
1. Label each event as "independent", "cause", "effect"
or "causal":
- "cause": Directly leads to or influences other events
- "effect": Results from or is influenced by other events
- "causal": Both causes and effects (two-way relationships)
- "independent": No causal relationships with other events
2. Look for causal indicators
3. Track relationships using event IDs
4. Enforce temporal constraints:
- cause_event can ONLY reference prior event IDs
- effect_event can ONLY reference later event IDs
5. Do not add more than {N} events
6. Output ONLY the JSON object, no additional text
Provide only the JSON object output.
```

C.2 VLM+LLM Validation Prompt

After LLM annotation, InternVL3.5-241B [13] generates captions from merged video clips for each proposed causal pair $(\mathcal{C}, \mathcal{E})$. The following prompt then instructs Llama-3.3-70B to score the causal support across four dimensions (Tier 2 validation, Section 3.2 of main paper). Critically, this validation task is far simpler than full causal graph prediction: the VLM captions only the *two* specific merged clips, not all $\binom{N}{2}$ pairs; this targeted grounding is what makes the validation reliable despite VLMs' known failure on full-graph causal prediction (Section 3.3 of main paper).

```
You are evaluating whether a proposed causal
relationship is supported by visual evidence in a video.
Given:
- Cause events: {cause_descriptions}
- Effect events: {effect_descriptions}
- Generated video caption: {generated_caption}

Evaluate the following four aspects (score 0.0-1.0 for
each):
1. PLAUSIBILITY (0.0-1.0):
Does the generated caption make the proposed causal link
plausible?
- 0.0: Caption contradicts the causal claim
- 0.5: Caption is neutral, neither supports nor contradicts
- 1.0: Caption strongly supports the causal claim
2. VISUAL_EVIDENCE (0.0-1.0):
How much explicit or implicit visual evidence for
causality appears?
- 0.0: No evidence of cause or effect visible
- 0.5: Both cause and effect visible, but connection unclear
```

```
- 1.0: Clear visual evidence of causal mechanism
3. STRENGTH (0.0-1.0):
How direct and strong is the causal connection?
Consider necessity (effect requires cause) and
sufficiency (cause typically produces effect).
- 0.0: Events appear independent
- 0.5: Weak or indirect connection
- 1.0: Strong, direct causal relationship
4. TRANSITION (0.0-1.0):
How smooth and coherent is the temporal
progression from cause to effect?
- 0.0: Abrupt or illogical transition
- 0.5: Acceptable temporal flow
- 1.0: Smooth, natural progression
Guidelines:
- Base judgment ONLY on provided information
- Do not introduce external knowledge
- Consider both what is shown and what is implied
- Be objective and consistent

Output JSON format:
{
    "plausibility": float,
    "visual_evidence": float,
    "strength": float,
    "transition": float,
    "overall": float (average of four scores),
    "reasoning":
        "brief explanation (max 50 words)"
}

Provide only the JSON object output.
```

D Additional Ablation Studies

This section supplements Section 4.4 (*Ablation Studies*) of the main paper. Table 4 in the main paper sweeps architectural and training design choices at a coarse level; the following tables provide the fine-grained hyperparameter sweeps that justify the specific values used throughout the paper.

D.1 DINOv2 Scene Detection Threshold

The DINOv2 frame sampling procedure (Section 3.3, *DINOv2 Frame Sampling* of main paper) identifies scene boundaries where cosine similarity between consecutive frame features falls below threshold θ. Table 9 justifies $\theta = 0.8$ cited in the main paper (Eq. 1 of main paper). Too low a threshold over-segments videos into many tiny, uninformative scenes; too high a threshold merges semantically distinct scenes, losing the temporal structure necessary for causal learning.

Table 9: DINOv2 scene-boundary threshold ablation (ActivityNet). Supports $\theta = 0.8$ used throughout the main paper (Eq. 1); $\theta = 0.8$ achieves the best F1.

θ	**F1**
0.50	0.70
0.60	0.69
0.70	0.71
0.80	**0.73**
0.90	0.70

D.2 Loss Balance Weight λ

The joint training objective (Eq. 14 of main paper, $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{caption}} + \lambda \cdot \mathcal{L}_{\text{causal}}$) requires balancing captioning and causal prediction losses. Table 10 justifies the choice $\lambda = 2.0$ cited in the main paper.

Table 10: Loss balance weight λ ablation (Eq. 14 of main paper, ActivityNet).

λ	F1	CIDEr
0.5	0.58	50.2
1.0	0.66	50.8
2.0	**0.73**	**52.0**
5.0	0.71	50.1
10.0	0.69	47.8

D.3 Hard Negative Mining Ratio

The hard negative mining mask M_{ij} (Eq. 16 of this appendix; Section 3.3 of main paper) retains the top-$k\%$ of most-confusing negatives. Table 11 justifies the 20% choice cited in the main paper.

Table 11: Hard negative mining ratio ablation (Section 3.3 of main paper, ActivityNet). 20% achieves the best F1.

HNM Ratio	F1
5%	0.68
10%	0.70
20%	**0.73**
30%	0.72
40%	0.71

D.4 False Negative Penalty Weight

The asymmetric penalty λ_{fn} (Eq. 17 of this appendix; Section 3.3 of main paper) amplifies the loss for false negatives; true causal links that the model fails to predict relative to false positives. Table 12 justifies $\lambda_{\text{fn}} = 10$ cited in the main paper.

Table 12: False negative penalty weight ablation (Section 3.3 of main paper, ActivityNet).

λ_{fn}	F1
1 (no penalty)	0.57
5	0.73
10	**0.73**
15	0.72
20	0.71

E Detailed Comparison with MECD

This section supplements Section 4.2 (*Main Results*) and the related work discussion of MECD [8, 9] in Section 2 of the main paper. We provide a full task-formulation comparison, class-imbalance analysis, and oracle performance breakdown that are summarised in one sentence in the main paper (Section 4.2: "0.74 F1 vs MECD's 0.71 on the MECD benchmark").

E.1 Task Formulation

MECD [8]: given N pre-segmented events (each with a video clip and textual caption provided by the dataset), predict which events $\{e_1, \ldots, e_{N-1}\}$ cause the final event e_N, giving $N - 1$ binary predictions, requiring human-annotated input.

Ours: given a raw video, generate N events with timestamps and captions end-to-end, then predict causal relationships for all $\binom{N}{2}$ ordered pairs. For a 10-event video, MECD makes 9 predictions (human input required); our model makes 45 from raw video alone.

E.2 Class Imbalance, Architecture, and Dataset Scale

Table 13 summarises key differences. The substantially higher imbalance ratio in our task (9:1–75:1 vs MECD's 2:1–4:1) explains why MECD's fixed positive weighting ($\lambda = 4.0$) is insufficient for our setting and necessitates our adaptive uncapped scheme (Appendix A.3, Eq. 15).

Table 13: MECD vs Causal-VLM task and design comparison. Our benchmark is $16\times$ larger; our task predicts all $\binom{N}{2}$ pairs rather than only the $N-1$ causes of the final event.

Aspect	MECD	Ours
Input	Clips + Captions	Raw video
Predictions	$N-1$	$\binom{N}{2}$
Typical class imbalance	2:1–4:1	9:1–75:1
Videos	1,105	21,586
Annotation LLM	GPT-4	Llama-3.3-70B

E.3 Performance Comparison

Table 14 details the two evaluation settings referenced briefly in Section 4.2 of the main paper. The *oracle* setting provides ground-truth events and captions directly to our Stage 2 causal head, isolating its reasoning quality under MECD-equivalent input conditions. The *full pipeline* evaluates our end-to-end model on raw video. The oracle gap (0.82 vs 0.74 final-event F1) reveals Stage 1 error propagation from imperfect temporal localization, consistent with the oracle ablation in Table 4f of the main paper (0.79 vs 0.73 full-graph F1 on ActivityNet).

Table 14: Performance comparison with MECD (Section 4.2 of main paper). Final-event F1 is extracted from our adjacency matrix $\mathbf{A}$ by selecting pairs (e_i, e_N), matching MECD's evaluation protocol. Our full pipeline outperforms MECD on final-event F1 (0.74 vs 0.71) while simultaneously solving the harder full-graph prediction problem.

Method	Final-Event F1	Full-Graph F1
MECD [8]	0.71	–
Ours (Oracle - GT events/captions)	**0.82**	**0.79**
Ours (Full pipeline - raw video)	0.74	0.73

F NExT-QA Transfer Evaluation

This section supplements the downstream transfer result mentioned in Section 4.2 of the main paper: "+6.4 on NExT-QA causal-Q accuracy." We provide the full breakdown by question type and comparison with state-of-the-art VideoQA methods to contextualise this transfer gain.

Setup. NExT-QA [4] is a video QA benchmark with 5,440 videos and ∼52K multiple-choice questions categorised into Causal (C, 48%), Temporal (T, 29%), and Descriptive (D, 23%) types. We use our Causal-VLM's VLM as a base VideoQA model, measuring how explicit causal supervision from DCC pre-training transfers to the causal subset of NExT-QA. Table 15 compares against state-of-the-art methods on the validation split (multiple-choice).

Table 15: NExT-QA validation set (multiple-choice): Causal (C), Temporal (T), Descriptive (D), and Overall (Acc) accuracy. Our causal features provide the largest gain on the Causal subset, with smaller gains on Temporal and Descriptive. Comparison methods are listed in order of Causal accuracy. †: zero-shot.

Method	Acc@C	Acc@T	Acc@D
SGVLM [61]	75.2	66.3	83.4
SeViLA [62]	74.2	69.4	81.3
BLIP-2 [10]	72.9	68.1	81.2
MoReVQA [63] †	70.2	64.6	–
HiTeA [64]	62.4	58.3	75.6
Causal-VLM (Ours)	***81.6***	***69.9***	*83.3*

G Blind Full-Graph Human Audit

This section gives the protocol and results of the audit summarised in Section 3.2.2 (step 4) of the main paper. The earlier human study (step 3) rates pairs that the LLM proposed as causal, which shows that included edges are plausible but cannot measure edges the LLM missed or the accuracy of its negative labels. The audit is designed to measure both.

Protocol. We sample 200 videos, 100 each from ActivityNet and YouCook2, stratified by event count, giving 2,008 eligible event pairs. Human annotators watch each video and label every eligible pair as

an edge or a non-edge using the causal-edge criteria in Section 3.2. They see no LLM output, neither the proposed labels. The released DCC labels are then scored against the resulting human reference.
Results. Table 16 reports the audit alongside the earlier validation study. In the earlier study, annotators see the proposed pair but not the LLM's confidence or reasoning.

Table 16: Human validation of DCC labels. The blind audit labels the full adjacency matrix and is independent of all the LLM output.

Study	Measure	Value
Blind audit (200 videos, 2,008 pairs)	Edge precision of released labels	0.92
	Edge recall of released labels	0.82
	Human agreement on non-edges (κ)	0.79
	Graph-level inter-annotator agreement	0.83
Pair study (2,000 videos, 9,574 pairs)	Human–human ICC	0.85
	Human–LLM agreement (κ)	0.76
	Llama–DeepSeek-R1 agreement (r)	0.90

Effect of the validation stage. We store the raw Llama-3.3-70B output before any validation or filtering. Scored against the released labels it reaches 0.824 F1 on ActivityNet and 0.733 on YouCook2, which quantifies the disagreement between the raw annotation output and the labels after validation. These F1 values do not specify the fraction of edges changed.
Scope. The audit estimates label quality on a stratified sample. It does not certify every graph in the 21.6K videos.

H Caption-to-LLM Baselines

This section expands the caption-to-LLM comparison in Section 4.2. Each LLM receives captions and timestamps with the DCC annotation prompt (Appendix C.1) and predicts the graph only, so the LLMs and our head use the same caption and timestamp inputs. Predicted timestamps remain subject to the event-matching protocol in the main paper.

Table 17: Caption-to-LLM baselines (causal F1). Top: Stage-1 predicted captions and timestamps, the input available at inference time. Bottom: ground-truth captions and timestamps (oracle). The LLMs emit discrete edge lists without per-pair confidence scores, so we report F1 rather than a ranked precision–recall curve for these models. Our head has AUPRC 0.687 (AN) and 0.616 (YC2).

Input to graph prediction	Model	ActivityNet	YouCook2
Predicted captions and timestamps	Llama-3.2-1B	0.204	0.207
	Llama-3.1-8B	0.271	0.221
	Llama-3.3-70B	0.487	0.369
	Causal-VLM head (11.6M)	**0.73**	**0.64**
Ground-truth captions and timestamps	Llama-3.3-70B, stored output	0.824	0.733
	Llama-3.3-70B, fresh run	0.822	0.730
	Causal-VLM head	0.79	0.74

On predicted input, Causal-VLM leads the best LLM by 0.24 F1 on ActivityNet and 0.27 on YouCook2. Inference latency is 0.7 ms per video for our head and 14.53 s per video for the 70B model. The ground-truth-caption rows are not a baseline for the task, because a system running on a new video has no human-written captions or timestamps. They also favour the LLM, since the released labels descend from the same model and prompt. The gap between the two input settings is informative: the LLM pipeline loses 0.34 (AN) and 0.36 (YC2) F1 when captions and timestamps must be predicted, while our head loses 0.06 and 0.10. This is consistent with the visual shuffling result in Table 4h.

I Additional Evaluation Metrics

This section supplements the causal F1 in Section 4.2 with threshold-free, per-video and calibration metrics, and with breakdowns of F1. All values are for Causal-VLM (full pipeline).

Table 18: Additional causal metrics. Headline F1 and CIDEr are means over 3 seeds (std $\pm$0.010 F1, $\pm$0.8 CIDEr).

Metric	ActivityNet	YouCook2
Micro-F1	0.73	0.64
AUPRC	0.687	0.616
Per-video macro-F1	within 0.04 of micro-F1	

Calibration and breakdowns. The calibration analysis gives an expected calibration error of 0.09. F1 on adjacent event pairs exceeds F1 on distant pairs by 0.13, and F1 on chains of four or more hops is 0.60.

J Robustness and Attribution

Causal loss versus added capacity. Table 19 trains the identical full architecture, with the same parameter count, epochs and schedule, with and without the causal loss. With the loss disabled, CIDEr moves by -1.5 (AN) and $+1.8$ (YC2) relative to Stage 1, against a cross-seed deviation of $\pm$0.8. With the loss enabled it rises by 7.8 and 5.5 over the $\lambda = 0$ model, in the same direction on both datasets.

Table 19: Attribution of caption gains (CIDEr).

Setting	ActivityNet	YouCook2
Stage-1 only	45.7	38.8
Full architecture, $\lambda = 0$	44.2	40.6
Full architecture, full loss	**52.0**	**46.1**

Caption errors and grounding. On ActivityNet, corrupting 25% and 50% of Stage-1 captions gives 0.69 and 0.55 F1, between the oracle (0.79), the predicted-caption pipeline (0.73) and the visual-only floor (0.34). Permuting the caption-to-event assignment within each video gives 0.55, and shuffling visual features across videos gives 0.52 (Table 4h).

Multi-domain training. Training jointly on ActivityNet and YouCook2 gives 0.60 F1 on YouCook2, against 0.47 for the ActivityNet-only model and 0.64 for the YouCook2-only model. Joint training recovers most of the transfer gap without closing it, which indicates that the two domains transfer substantially while carrying partly distinct causal structure.

K Qualitative Results

This section provides additional qualitative examples supplementing Figure 4 in the main paper (Section 4 of main paper). Figure 5 shows eight examples spanning all four causal pattern types identified in the DCC benchmark: linear chains, parallel patterns, complex graphs, and long sequential chains. Green checkmarks indicate correct predictions; red crosses indicate errors; circle-slash indicates invalid structured output from a baseline model.

(a)–(c) Linear chains (3 events). Our model correctly predicts $E_1 \to E_2 \to E_3$ in all three cases. Baselines either skip the intermediate link (predicting $E_1 \to E_3$ directly) or produce invalid structured output for middle-chain events, consistent with the main paper's finding that baselines fail to reason about indirect causal relationships.



Figure 5: Qualitative predictions across all causal pattern types (extends Figure 4 in the main paper). Each row pair shows event frames with captions and the predicted adjacency matrix. Rows = causes; columns = effects. **Green** ✓ = correct; **red** × = error; ⊘ = invalid structured output (model failed to produce valid JSON). Our model (Causal-VLM) outperforms fine-tuned PLM-3B [12] and InternVL-4B [13] on all pattern types.

(d)–(e) Parallel patterns. In example (e), E_5 (drizzle dressing on salad) requires *both* E_3 (toss) and E_4 (combine molasses), a parallel dependency that temporal-ordering models cannot capture. Our model identifies both $E_3 \to E_5$ and $E_4 \to E_5$; both baselines predict only the simpler $E_1 \to E_2$ link.

(f) Complex graph (7 events). This video contains mixed independent, causal, and parallel relationships. Our model correctly predicts 4 of 5 causal links; baselines predict only 1. The one incorrect prediction (over-predicting E_7 as causal when it is independent) reflects the over-prediction failure mode identified in the main paper (Section 4.4).
(g)–(h) Long chains (8 events). Our model predicts 6 of 7 causal links in both examples. PLM-3B predicts 2–3; InternVL-4B predicts 3–4. The incorrect prediction in each case involves temporally adjacent events with visually similar content, consistent with failure mode (1) in Section 4.4 of the main paper.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction state three claims that are directly supported by experiments in the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The Limitation section is in the main paper.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [N/A]
   Justification: The paper makes no formal theoretical claims. All contributions are empirical: a benchmark dataset, an architecture, and experimental results.
4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: Architecture is fully described in Section 3.3 with all equations and dimensions (Appendix A, Table 7). Training hyperparameters for both stages are provided in Appendix A (Tables 5–6). Dataset splits are specified in Appendix B. Annotation prompts are in Appendix C. Code and dataset are publicly released.
5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: The DCC benchmark dataset is released at `https://huggingface.co/datasets/causalvlm/DCC` under CC BY-NC-ND 4.0, with a Croissant metadata file submitted alongside this paper. The Causal-VLM code is released at `https://github.com/causalvlm/CausalVLM`.
6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: Data splits are given in the Appendix B, hardware (NVIDIA H200 GPUs), batch size (16), and precision (bfloat16) are specified in Section 4.1. Full hyperparameter tables including learning rates per phase, warmup steps, loss balance weights, and hard negative mining ratios are in Appendix A (Tables 5–6). Hyperparameter selection methodology is described in Appendix D.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: Headline results are means over 3 seeds, with standard deviation ± 0.010 F1 and ± 0.8 CIDEr (Section 4.1 and Appendix I).
8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: Section 4.1 states that all models are implemented in PyTorch with bfloat16 precision on NVIDIA H200 GPUs, batch size 16. Appendix A provides epoch counts per training phase (Stage 1: 25 total epochs; Stage 2: 20 total epochs) and gradient accumulation factor (4), from which total compute can be estimated..

9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?
   Answer: [Yes]
   Justification: The paper uses publicly available datasets (ActivityNet, YouCook2) under their respective licenses. No personally identifiable information is collected or released. The dataset and model are released for non-commercial research use (CC BY-NC-ND 4.0), reducing misuse risk.
10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [Yes]
    Justification: The Conclusion discusses positive impacts: improved audio descriptions for visually impaired users (accessibility), robotic task planning with causal awareness, and more informative tutorial systems. Potential negative impacts are limited but acknowledged: (1) incorrect causal predictions could mislead automated systems in safety-critical robotics applications, mitigated by the model's high precision (0.77–0.90) and the oracle gap analysis showing a known error ceiling; (2) the dataset inherits any biases present in ActivityNet and YouCook2 (predominantly English-language Western cooking and sports activities). We release under a non-commercial license to limit deployment without additional safety review.
11. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?
    Answer: [N/A]
    Justification: The DCC dataset contains only structured causal annotations (event labels and adjacency matrices) over existing publicly licensed video datasets; it does not contain raw video content, personally identifiable information, or generated media. The Causal-VLM model predicts causal relationships in instructional videos and poses no meaningful dual-use risk. The CC BY-NC-ND 4.0 license prevents commercial exploitation.
12. **Licenses for existing assets**
    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
    Answer: [Yes]
    Justification: All assets are properly cited. ActivityNet Captions [16] is used under its Creative Commons license for research purposes. YouCook2 [17] is used under its BSD-style research license. Llama-3.3-70B and Llama-3.1-3B [18] are used under Meta's Llama Community License. DINOv2 [52] and the Perception Encoder [53] are used under Apache 2.0. InternVL3.5 [13] is used under Apache 2.0. EMScore [19] is used for evaluation under its MIT license.
13. **New assets**
    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
    Answer: [Yes]
    Justification: The DCC benchmark dataset is released on HuggingFace (`https://huggingface.co/datasets/causalvlm/DCC`) with: (1) a Croissant metadata file describing schema, splits, and provenance; (2) full annotation generation and validation prompts (Appendix C); (3) dataset statistics (Appendix B, Table 8).
14. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
    Answer: [Yes]
    Justification: Section 3.2 describes the human validation study. The annotation process and rating scale are described in Section 3.2.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
    Answer: [N/A]
    Justification: The human annotation task involved rating the causal plausibility of pairs of publicly available instructional video clips on a 5-point scale. No personally identifiable information was collected from annotators.
16. **Declaration of LLM usage**
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?
    Answer: [Yes]
    Justification: LLMs are a core, original component of this research and are fully declared. Llama-3.3-70B is used for causal annotation generation (Section 3.2, prompt in Appendix C); InternVL3.5-241B is used for VLM-based content verification (Section 3.2); Llama-3.3-70B is used for scoring the four validation dimensions (Section 3.2).