GENERATIVE CO-EVOLUTION: RETHINKING WEAKLY-SUPERVISED VIDEO HIGHLIGHT DETECTION IN THE MLLM ERA

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly-supervised video highlight detection (WSVHD) aims to localize highlight moments using only video-level labels. While Multimodal Large Language Models (MLLMs) have strong reasoning capabilities, adapting them to WSVHD remains underexplored and is severely bottlenecked by sparse supervision and complex topic semantics. Conventional supervised fine-tuning (SFT) struggles under this signal scarcity, whereas semi-supervised approaches are constrained by static unlabeled datasets and laborious manual curation that stall rapid model iteration. To address this problem, we introduce a novel generative-adversarial post-training framework that drives MLLM performance improvement for WSVHD without human-curated unlabeled data. Rather than isolated multiple instance learning, we introduce a generative model that co-evolves with the MLLM using an efficient video synthesis tool. Through parameter updates guided by Group Relative Policy Optimization (GRPO), the generative model moves beyond a static data generator and synthesizes realistic, topic-relevant training video features to increase the MLLM's discriminative capability for WSVHD. Extensive experiments demonstrate that our co-evolutionary paradigm significantly outperforms standard supervised fine-tuning and classical semi-supervised learning, establishing new state-of-the-art performance for WSVHD across three public benchmarks. Code will be made publicly available.

1 INTRODUCTION

The daily proliferation of video content—evidenced by over 500 hours of video uploaded to YouTube every minute[1]—has created a pressing need for automatic video understanding and editing. On the producing side, content creators require automated ways to extract highlight clips from raw footage. On the consuming side, viewers demand quick access to the most compelling moments of long-form media. To address this need, video highlight detection (VHD) delivers substantial commercial value through studying algorithms Sun et al. (2014); Moon et al. (2023); Xiao et al. (2024); Gordeev et al. (2026) that localize watch-worthy video segments without costly human review.

As a critical task in video understanding, learning-based VHD has been studied extensively over the past decade, including fully-supervised and weakly-supervised methods. The former rely on precise temporal boundary annotations, which are prohibitively expensive to collect and bottleneck rapid model iteration. To circumvent this, weakly-supervised video highlight detection (WSVHD) assumes the absence of frame-level annotations and mainly uses video-level labels. Leveraging breakthroughs in deep representation learning and Multiple Instance Learning (MIL) (Xiong et al., 2019; Hong et al., 2020; Li et al., 2024), WSVHD has achieved significant progress.

Recently, the emergence of multi-modal large language models (MLLMs) (Liu et al., 2023; Wu et al., 2023; Zhang et al., 2024) has revolutionized video understanding from signal-level perception to cognitive-level reasoning, with unprecedented semantic representational power emerging from large-scale data and huge model parameters. Various classical video understanding tasks are reformulated as post-training of an MLLM (Tang et al., 2025). Recent efforts have begun to explore

[1] https://www.youtube.com/howyoutubeworks/product-features/search/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

adapting MLLMs for VHD while relying on fully-supervised paradigms (Lee et al., 2025). This leaves a critical, underexplored question: *How to effectively post-train MLLMs for WSVHD?*

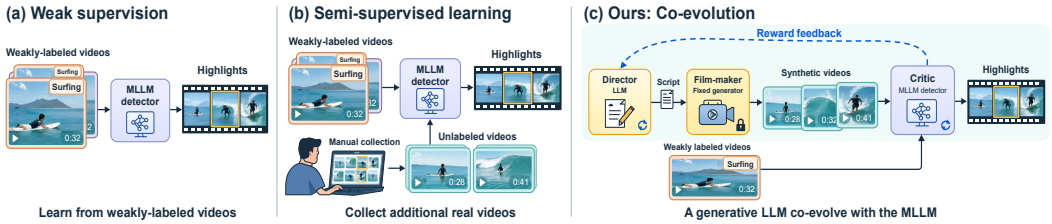

Figure 1: Diagrams of different post-training schemes for MLLM-based WSVHD: (a) fine-tuning a MLLM with weakly-labeled videos, (b) utilizing manually collected unlabeled data for semi-supervised learning, and (c) our generative co-evolution paradigm where two models interact and improve themselves.

In this work, we first apply off-the-shelf supervised fine-tuning (SFT) techniques for WSVHD, whose suboptimal results highlight the challenges of the low-data regime. We then propose a strong baseline that adapts an MLLM for WSVHD by utilizing MIL to train the task-specific adapter modules. Leveraging unlabelled video data available online, we take a further step by enhancing the MLLM's capabilities with semi-supervised learning (SSL). By incorporating the Mean Teacher (Tarvainen & Valpola, 2017) framework into training, we achieved additional improvements at the cost of slower model iteration due to the extra manual data collection stage.

To further mitigate this problem, we propose an innovative training framework featuring co-evolution of a discriminative MLLM and a paired generative LLM for WSVHD as shown in Fig. 1. Our approach removes the need for real unlabeled data collection for model regularization through a novel incorporation of efficient video synthesis into automatic model improvement. Rather than training the MLLM in isolation, we introduce an adversarial framework featuring a "director" LLM equipped with a video generation tool. This director synthesizes novel video content to regularize the MLLM post-training, while learning to generate topic-relevant, realistic highlights. By alternating between adapting the LLM and the MLLM, we strengthen the LLM for highlight generation and the MLLM for discrimination without manual data collection. Extensive experiments show that this co-evolution framework significantly outperforms standard SFT and classical SSL for WSVHD and achieves state-of-the-art results on public benchmarks.

Our main contributions are summarized as follows:

- We systematically investigate the post-training of MLLMs dedicated to WSVHD.
- We propose a novel post-training scheme that leverages an opponent generative model to drive MLLM co-evolution with synthetic data. We demonstrate that this paradigm yields substantial performance gains over traditional SFT and SSL baselines and enables model self-improvement without manual data collection.
- Our proposed method achieves new state-of-the-art performance across three public benchmark datasets.

2 Related Work

Weakly-Supervised Video Highlight Detection. Prior WSVHD studies focus on learning segment-level representations from video-level supervision using duration cues (Xiong et al., 2019), contrastive learning (Hong et al., 2020), global context (Ye et al., 2021), or cross-modal alignment and masked reconstruction (Li et al., 2024). These pre-MLLM methods train small-scale models from scratch using raw audio-visual features. In contrast, we leverage the rich vision-language knowledge in modern MLLMs and adapt them to this challenging task with a new training paradigm featuring a co-evolving video generation model.

MLLMs for Video Understanding. MLLMs have extended the capabilities of LLMs (Touvron et al., 2023; Wei et al., 2022) to non-textual modalities and have become increasingly popular in recent video understanding architectures (Lin et al., 2024). While recent works like LLMVS (Lee et al., 2025) apply LLMs to fully-supervised VHD, our study provides an in-depth exploration of adapting MLLMs for the significantly more challenging weakly-supervised problem.

Post-Training of MLLMs. MLLMs are typically adapted to downstream tasks via SFT (Ouyang et al., 2022) or RLHF (Rafailov et al., 2023), often utilizing parameter-efficient techniques (Hu et al., 2022). Recently, temp-R1 (Wu et al., 2026) and time-R1 (Wang et al., 2026b) applied GRPO (Shao et al., 2024) for video understanding. Building on this momentum, we propose a novel post-training framework tailored for WSVHD, which not only demonstrates superior performance over classical SFT, but also enjoys extra model regularization without manual data curation.

Semi-Supervised Learning with Foundation Models. Recent literature actively explores optimal strategies for utilizing unlabeled data to adapt highly parameterized foundation models (Zhang et al., 2026). Similarly, we strive to extract supplementary supervisory signals in low-data regimes. Uniquely, instead of maintaining a static repository of unlabeled videos, we integrate an automated video synthesis mechanism directly into the model adaptation loop. This fundamentally bypasses the laborious curation, pre-processing, and inherent biases in real-world datasets.

Self-Improvement in Large Models. To bypass expensive human annotations required by SFT and traditional RLHF, recent studies explore model self-evolution (Yuan et al., 2024). For instance, Vision-Zero (Wang et al., 2026a) enables MLLMs to iteratively improve through a multi-agent zero-sum game. Inspired by this, we propose a framework featuring the co-evolution of a generative LLM and a discriminative MLLM. This framework yields substantial model improvements with negligible human oversight, marking the first successful attempt at training an MLLM from labor-free model interaction for WSVHD.

3 METHODS

3.1 PROBLEM FORMULATION

Weakly-supervised video highlight detection aims to infer segment-level highlight scores with video-level annotations. A simplified analogy is predicting which key in a chain of keys opens a door when you only know that some key chains open a door. This abstraction was described in the problem of predicting drug molecule activity addressed in the 1990s using the MIL framework (Dietterich et al., 1997), where bag-level supervision guides instance-level predictions.

Formally, we define a dataset of N videos as $\mathbb{D} = \{(\mathbb{V}_i, y_i)\}_{i=1}^{N}$. Each video $\mathbb{V}_i$ acts as a bag of T_i temporal segments, represented by instance feature matrices $\boldsymbol{V}_i = [\boldsymbol{v}_{i,1}, \boldsymbol{v}_{i,2}, \ldots, \boldsymbol{v}_{i,T_i}] \in \mathbb{R}^{d \times T_i}$, where $\boldsymbol{v}_{i,t} \in \mathbb{R}^d$ is the feature vector of the t-th segment. The scalar $y_i \in \{0, 1\}$ denotes the video-level label, indicating the presence ($y_i = 1$) or absence ($y_i = 0$) of highlight moments in the video. Our goal is to learn a scoring function $h : \mathbb{R}^{d \times T_i} \to [0, 1]^{T_i}$ that maps the video instance features to a temporal score vector $\boldsymbol{s}_i = h(\boldsymbol{V}_i) = [s_{i,1}, s_{i,2}, \ldots, s_{i,T_i}]^\top \in [0, 1]^{T_i}$, without access to segment-level ground-truth labels. Here $s_{i,t} \in [0, 1]$ denotes the highlight score of the t-th segment.

3.2 A BASELINE: MLLM ADAPTATION WITH MULTIPLE INSTANCE LEARNING

We started by treating WSVHD as a video question-answering problem and used an SFT baseline that fine-tunes an MLLM with video-level answers and asks it to produce clip-level highlight scores. This baseline is described in A.2 and performs poorly, highlighting the challenge of lacking clip-level supervision. To establish a stronger baseline for WSVHD, we adapt an MLLM within a MIL-guided framework as shown in Fig. 2. We freeze the MLLM backbone layers and add lightweight, learnable prediction modules. Given a video $\mathbb{V}_i$ with T_i temporal segments, we first extract unimodal visual and audio feature matrices $\boldsymbol{V}_i^{(v)}, \boldsymbol{V}_i^{(a)} \in \mathbb{R}^{d \times T_i}$ using the pre-trained Qwen-3 MLLM encoders (Bai et al., 2025; Shi et al., 2026). To model temporal dependencies, these features are processed through modality-specific self-attention blocks, yielding $\tilde{\boldsymbol{V}}_i^{(v)} = \text{SelfAttn}_v(\boldsymbol{V}_i^{(v)})$ and $\tilde{\boldsymbol{V}}_i^{(a)} = \text{SelfAttn}_a(\boldsymbol{V}_i^{(a)})$. For cross-modal fusion, we employ a multi-head cross-attention layer where queries are derived from the visual features, and keys/values

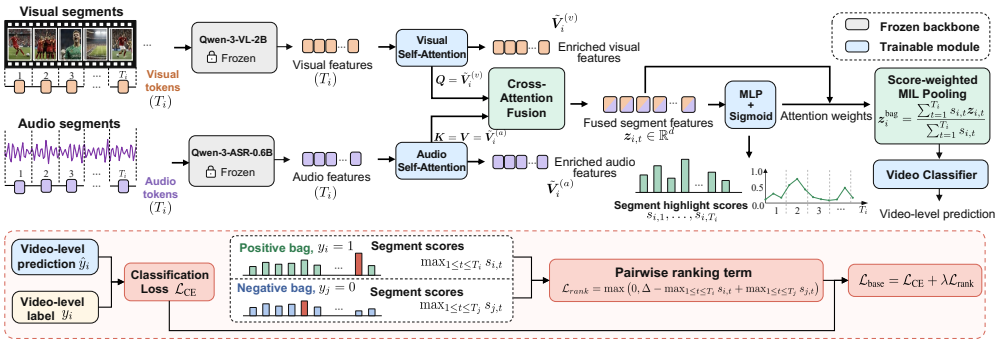

Figure 2: Diagram of the MIL baseline that extracts visual-audio features, followed by modal-specific self-attention and cross-attention fusion layers. A highlight prediction head outputs video-level and segment-level highlight scores where a bag-level classification loss and a ranking loss is used for model supervision.

are derived from the audio features. This produces the multimodal contextualized representation matrix $\boldsymbol{Z}_i = [\boldsymbol{z}_{i,1}, \boldsymbol{z}_{i,2}, \ldots, \boldsymbol{z}_{i,T_i}] \in \mathbb{R}^{d \times T_i}$, where $\boldsymbol{z}_{i,t} \in \mathbb{R}^d$ denotes the feature vector for the t-th segment.

These contextualized features are fed into a Multi-Layer Perceptron (MLP) with a sigmoid activation $\sigma(\cdot)$ to compute the instance-level highlight scores $s_{i,t} = \sigma(\text{MLP}_{\text{ins}}(\boldsymbol{z}_{i,t})) \in [0, 1]$. Using $s_{i,t}$ as temporal attention weights, we aggregate the segment representations into a video-level bag feature vector $\boldsymbol{z}_i^{\text{bag}} = \sum_{t=1}^{T_i} s_{i,t}\boldsymbol{z}_{i,t} / \sum_{t=1}^{T_i} s_{i,t}$. A video-level classification head then estimates the probability of video $\mathbb{V}_i$ containing a highlight moment as $\hat{y}_i = \sigma(\text{MLP}_{\text{bag}}(\boldsymbol{z}_i^{\text{bag}}))$.

The model parameters are optimized end-to-end using a composite MIL loss (Hong et al., 2020; Li et al., 2024). First, video-level classification is supervised via a binary cross-entropy loss over a mini-batch $\mathbb{B}$, defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathbb{B}|} \sum_{i \in \mathbb{B}} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \tag{1}$$

Secondly, a pairwise ranking loss is applied between positive videos $\mathbb{B}^+ = \{i \in \mathbb{B} \mid y_i = 1\}$ and negative videos $\mathbb{B}^- = \{j \in \mathbb{B} \mid y_j = 0\}$. This auxiliary loss ensures that the peak segment score in any positive video exceeds that of a negative video by at least a predefined margin Δ as:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathbb{B}^+||\mathbb{B}^-|} \sum_{i \in \mathbb{B}^+} \sum_{j \in \mathbb{B}^-} \max(0, \Delta - \max_t s_{i,t} + \max_t s_{j,t}) \tag{2}$$

The training objective is minimized as $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{rank}}$, where λ is a balancing hyperparameter.

3.3 Regularized highlight score prediction using unlabeled videos

To mitigate the supervision scarcity in our baseline, we incorporate a semi-supervised mean teacher framework (Tarvainen & Valpola, 2017) leveraging curated unlabeled videos. The architecture consists of an active student model (parameters $\boldsymbol{\theta}^S$) and a structurally identical teacher model (parameters $\boldsymbol{\theta}^T$). The teacher is updated via an exponential moving average (EMA) of the student at each training step: $\boldsymbol{\theta}^T \leftarrow \alpha\boldsymbol{\theta}^T + (1 - \alpha)\boldsymbol{\theta}^S$, where $\alpha \in [0, 1)$ is the momentum hyperparameter.

For a mini-batch of unlabeled videos $\mathbb{B}_u$, we enforce prediction consistency between the networks. Given a video $\mathbb{V}_i \in \mathbb{B}_u$, let $\boldsymbol{s}_i^S \in [0, 1]^{T_i}$ and $\hat{y}_i^S \in [0, 1]$ represent the student's instance-level temporal score vector and bag-level highlight probability, while $\boldsymbol{s}_i^T$ and $\hat{y}_i^T$ denote the teacher's corresponding outputs. We align the segment-level predictions using Mean Squared Error (MSE) and the global bag-level probability distributions using Kullback-Leibler (KL) divergence. The overall consistency loss is formulated inline as $\mathcal{L}_{\text{MT}} = \frac{1}{|\mathbb{B}_u|} \sum_{i \in \mathbb{B}_u} (\frac{1}{T_i} ||\boldsymbol{s}_i^S - \boldsymbol{s}_i^T||^2 + \beta D_{\text{KL}}(\boldsymbol{p}_i^T \| \boldsymbol{p}_i^S))$, where β balances the bag-level alignment. This regularization encourages stable predictions on unseen data and is directly added to the baseline objective $\mathcal{L}_{\text{base}}$.

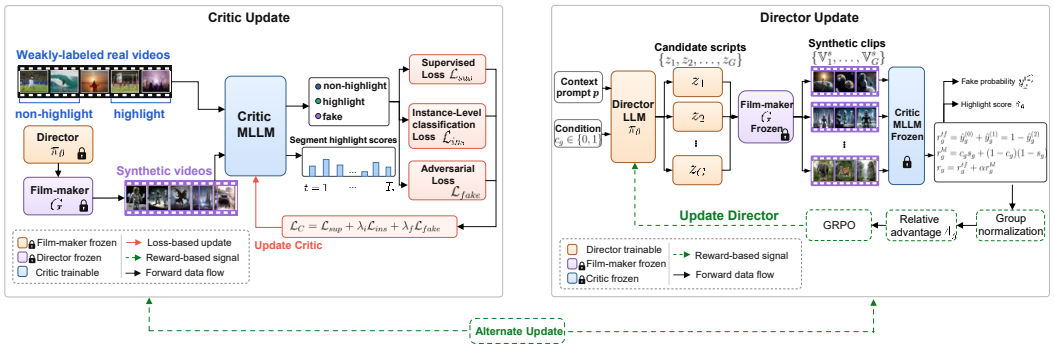

Figure 3: The proposed generative co-evolution post-training paradigm. The critic updates with additional synthetic videos generated from the director. The Director updates with the critic's reward.

3.4 Toward Labor-less Model Iteration via Generative Co-evolution

The previous semi-supervised learning attempt can improve model performance without extra frame-level annotations for WSVHD. However, it faces several fundamental bottlenecks:

Requirement for manual data curation: Manual search, quality assurance, and pre-processing of unlabeled data are prohibitively time-consuming and stall rapid model iteration.

Inflexibility to varying highlight topics: A static repository of unlabeled videos struggles to scale to the dynamic, ever-changing highlight topics that emerge in user-generated content communities.

Vulnerability to dataset bias and label noise: Real-world data harbors inherent biases that induce shortcut learning. In addition, a single video often contains content corresponding to multiple topics, which manual curation frequently overlooks. Consequently, during topic-specific MIL, presumed negative videos may inadvertently contain positive instances, introducing severe label noise that hinders training.

Instead of laboriously hunting for domain-specific unlabeled videos, we ask: *can we automatically synthesize the targeted video data we need, and moving beyond a mere static tool, can we drive the generator to self-improve through deep synergy with our existing MIL framework?* Recent breakthroughs in model self-evolution (Wang et al., 2026a) suggest a paradigm shift. Unlike SFT bounded by fixed datasets, or traditional Reinforcement Learning from Human Feedback (RLHF) relying on static environmental rewards, enhanced model capabilities can be achieved dynamically through interactions among models. Motivated by this, alongside recent progress in text-to-video generation (Wan et al., 2025; Zhang et al., 2025), we propose a *generative co-evolution* training paradigm as shown in Fig. 3.

Our framework comprises three components: a *Director* (an LLM), a *Film-maker* (a text-to-video synthesis tool), and a *Critic* (the target MLLM optimizing for WSVHD). The Director, parameterized by vector $\boldsymbol{\theta}$ and denoted as $\pi_{\boldsymbol{\theta}}(z \mid p)$, receives a context prompt p and auto-regressively generates a detailed scene script z. The Film-maker, $G(z)$, renders this script into a synthetic video clip. By sampling various prompts, we construct a dynamic dataset of synthetic videos, $\mathbb{D}_s = \{(\mathbb{V}_k^s, c_k)\}_{k=1}^K$, where $c_k \in \{0, 1\}$ serves as the conditional generation prompt label (highlight or non-highlight). The Critic evaluates both these synthetic clips and the weakly-labeled real videos. We alternate the optimization of the Director and the Critic.

Director Update. We optimize the Director $\pi_{\boldsymbol{\theta}}$ using Group Relative Policy Optimization (GRPO) while freezing the Critic's parameters. Given a prompt p, the Director generates a group of G candidate scripts $\{z_1, z_2, \ldots, z_G\}$, which the Film-maker renders into synthetic clips $\{\mathbb{V}_1^s, \ldots, \mathbb{V}_G^s\}$. The Critic evaluates these clips, returning a composite reward r_g for each sample.

Bag-Level True-Fake Reward. Borrowing principles from semi-supervised Generative Adversarial Networks (Dai et al., 2017), we incentivize the Director to synthesize realistic videos capable of

deceiving the Critic. We expand the Critic's video-level classification head to output a probability distribution over three distinct classes: $c \in \{0, 1, 2\}$, denoting *non-highlight*, *highlight*, and *fake*, respectively. The synthetic clips within a group act as a video bag for this classification. The base reward measures the bag's success at masquerading as real content (i.e., dodging the *fake* label), computed as $r_g^{tf} = \hat{y}_g^{(0)} + \hat{y}_g^{(1)} = 1 - \hat{y}_g^{(2)}$.

Instance-Level Highlight Reward. Beyond realism, we explicitly push the Director to generate prototypical highlight and non-highlight clips to strengthen the supervisory signal. We use the prompt condition $c_g \in \{0, 1\}$, where $c_g = 1$ targets a highlight moment and $c_g = 0$ targets background context. s_g is the instance-level highlight score predicted by the Critic. We define the instance-level reward as $r_g^{hl} = c_g s_g + (1 - c_g)(1 - s_g)$.

The total reward for the g-th clip linearly combines realism and semantic accuracy as $r_g = r_g^{tf} + \alpha r_g^{hl}$. We normalize these rewards within the group to calculate the relative advantage $A_g = (r_g - \mu_r)/\sigma_r$, where μ_r and σ_r are the group mean and standard deviation. The Director is updated by maximizing the standard GRPO surrogate objective:

$$\mathcal{J}_D(\boldsymbol{\theta}) = \mathbb{E}_{p,\{z_g\}_{g=1}^{G}} \left[\frac{1}{G} \sum_{g=1}^{G} \min \left(\frac{\pi_{\boldsymbol{\theta}}(z_g \mid p)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(z_g \mid p)} A_g, \text{clip} \left(\frac{\pi_{\boldsymbol{\theta}}(z_g \mid p)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(z_g \mid p)}, 1 - \epsilon, 1 + \epsilon \right) A_g \right) \right] \quad (3)$$

Critic Update. During its optimization phase, the Critic abandons unlabeled real data and solely processes two parallel data streams: real weakly-labeled bags $\mathbb{D} = \{(\mathbb{V}_i, y_i)\}_{i=1}^{N}$ and the generated synthetic bags $\mathbb{D}_s = \{(\mathbb{V}_k^s, \boldsymbol{c}_k)\}_{k=1}^{K}$. For an input video, the Critic extracts the aggregated bag-level feature vector $\boldsymbol{z}^{\text{bag}}$ and outputs a class probability vector $\hat{\boldsymbol{y}} = \text{Softmax}(\text{MLP}_{\text{vid}}(\boldsymbol{z}^{\text{bag}})) \in [0, 1]^3$. The Critic simultaneously optimizes the supervised loss for WSVHD and additional losses from the synthetic videos as the following three-part objective:

1. **Supervised Loss:** Enforces correct highlight classification on labeled data as $\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{base}}$.
2. **Adversarial Loss:** Trains the Critic to identify synthetic videos as fake (class 2), computed as $\mathcal{L}_{\text{fake}} = -\mathbb{E}_{(\mathbb{V}_k^s, \boldsymbol{c}_k) \sim \mathbb{D}_s} \left[\log \hat{y}_k^{(2)} \right]$.
3. **Instance-Level Classification Loss:** Leverages the prompt condition sequence $\boldsymbol{c}_k$ used to generate the synthetic video as segment-level pseudo-labels for the instance scores $s_{k,t}$, formulated as $\mathcal{L}_{\text{ins}} = -\mathbb{E}_{(\mathbb{V}_k^s, \boldsymbol{c}_k) \sim \mathbb{D}_s} \left[\frac{1}{T_k} \sum_{t=1}^{T_k} \left(c_{k,t} \log s_{k,t} + (1 - c_{k,t}) \log(1 - s_{k,t}) \right) \right]$.

The composite Critic loss is minimized as $\mathcal{L}_C = \mathcal{L}_{\text{sup}} + \lambda_f \mathcal{L}_{\text{fake}} + \lambda_i \mathcal{L}_{\text{ins}}$, governed by balancing hyperparameters λ_f and λ_i. We alternate the update of the Director and the Critic, and the training process is summarized in A.5.1.

Our generative co-evolution framework shares the adversarial spirit of semi-supervised Generative Adversarial Networks (GANs) (Dai et al., 2017) yet differs fundamentally. Because the discrete script generation and text-to-video rendering steps are non-differentiable, we formulate the interaction as a reinforcement learning problem driven by scalar rewards from the Critic. Building on prior works, our true-fake reward and adversarial loss inherit the theoretical advantage of refining the classifier's decision boundaries. However, rather than optimizing solely for data realism, our ultimate goal is to synthesize hard, semantically conditioned training samples. Tailored to the specific characteristics of WSVHD, we further introduce an instance-level reward and loss. This explicit formulation drives the Director to generate distinct highlight and non-highlight segments, thereby providing the Critic with richer, fine-grained supervision to actively boost its downstream performance.

4 EXPERIMENTS

To evaluate our generative co-evolution framework for WSVHD, we first conduct an extensive ablation study to validate our core design choices. We then compare our framework with previous studies across three standard public benchmarks. Implementation details and extended results are provided in the appendix.

4.1 Benchmarks and Evaluation Metrics

We evaluate our method on three datasets spanning diverse highlight domains: YouTube Highlights (Sun et al., 2014), TVSum (Song et al., 2015), and MoSu (Kim et al., 2026). The former two are established WSVHD benchmarks extensively used in prior work, and the recent MoSu dataset offers a significantly larger-scale evaluation.

TVSum contains 50 videos across 10 distinct topics. Following standard protocols (Cai et al., 2018; Xiong et al., 2019), we evaluate mAP using the evaluation toolkit from Hong et al. (2020), defining the top-scoring 50% of human-annotated shots as ground-truth highlights.

YouTube Highlights spans six domains: dogs, gymnastics, parkour, skating, skiing, and surfing. Each video contains segment-level binary highlight annotations. We follow the data splits established by Liu et al. (2022) and compute mAP using the tools from Hong et al. (2020).

MoSu (Most Replayed Multimodal Video Summarization) is a large-scale benchmark comprising 52,678 videos ($\sim$4,000 hours), with ground-truth highlight scores derived from viewer replay behavior. Videos are clustered into 10 coarse categories. To ensure our framework remains agnostic to specific visual/audio encoders, we bypass raw video processing and directly utilize the pre-extracted features provided by the authors.

Table 1: Summary of ablation studies on the TVSum dataset. **R.U.D.** denotes real unlabeled data, **Synth.** denotes synthetic data, **Clips** denotes the number of synthetic clips, **Rew.** denotes reward type (Bag = bag-only, Full = full reward), and **Dir. Up.** denotes director update. V and V+A denote visual-only and audio-visual inputs, respectively. The two SFT baselines are described in A.2.

Method	R.U.D.	Synth.	Clips	Rew.	Dir. Up.	Average Top-5 mAP
Training Paradigm						
SFT (Qwen3-VL-2B, V)	–	–	–	–	–	0.729
SFT (Qwen2.5-Omni-3B, V+A)	–	–	–	–	–	0.693
MLLM + MIL	–	–	–	–	–	0.768
MLLM + MIL + SSL	✓	–	–	–	–	0.773
Director Update						
Ours (no director update)	–	✓	100	–	✗	0.778
Ours (Bag-only Reward)	–	✓	100	Bag	✓	0.782
Number of Synthetic Clips						
Ours (Full, 100 Clips)	–	✓	100	Full	✓	0.800
Ours (Full, 300 Clips)	–	✓	300	Full	✓	0.802
Ours (Full, 500 Clips)	–	✓	500	Full	✓	0.788

4.2 Ablation Study

Table 1 ablates our training paradigm and architectural components on the TVSum dataset.

Efficacy of MIL Adaptation. Standard supervised fine-tuning (SFT) struggles to leverage weak supervision effectively. In contrast, incorporating MIL architectures with task-specific priors consistently improves both representation learning and overall detection accuracy.

Impact of Semi-Supervised Regularization. Applying highlight score regularization via auxiliary unlabeled data improves upon the MIL baseline, confirming that the MeanTeacher paradigm effectively mitigates data scarcity without additional frame-level annotations.

Superiority of Generative Co-Evolution. Our generative co-evolution paradigm achieves the highest accuracy. Notably, it surpasses the MeanTeacher approach (which uses real unlabeled data), demonstrating that targeted synthetic videos provide a stronger supervisory signal while eliminating manual curation bottlenecks.

Efficacy of Director Update. To validate co-evolution, we compare our model against a static baseline (frozen LLM, no GRPO). Tab. 1 shows static data yields a 0.778 mAP, while dynamically updating the Director via GRPO boosts it to 0.800 (100-clip setting). This confirms dynamic updates generate more challenging, topic-relevant synthetic samples to better sharpen the Critic's boundaries.

Influence of Reward Formulation. Removing instance-level highlight rewards yields a "Bag-only" variant (Dai et al., 2017). While this outperforms standard SSL, integrating instance-level rewards yields further gains by forcing the Director to capture fine-grained highlight semantics rather than just realistic background distributions. As shown in Fig. 4, the Director obtains increasing rewards with a healthy standard deviation during training, successfully generating videos that capture core highlight concepts.

Sensitivity to Number of Synthetic Clips. Evaluating synthetic clip numbers of 100, 300, and 500 reveals that 300 achieves optimal performance, striking a balance between real-data utilization and synthetic-data augmentation.

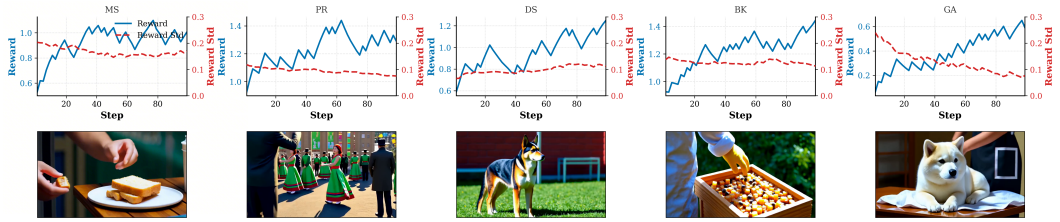

Figure 4: Visualization of the director rewards and synthetic videos during the training process on the TVSum dataset for topics MS, PR, DS, BK, and GA.

Table 2: Top 5 mAP on the TVSum dataset. Best and second-best are **boldfaced** and underlined, respectively as also in the following tables.

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Average
KVS (Potapov et al., 2014)	0.353	0.441	0.402	0.417	0.382	0.403	0.397	0.342	0.419	0.394	0.398
DPP (Gong et al., 2014)	0.399	0.453	0.457	0.462	0.437	0.446	0.442	0.395	0.464	0.449	0.447
sLSTM (Zhang et al., 2016)	0.411	0.462	0.463	0.477	0.448	0.461	0.452	0.406	0.471	0.455	0.451
SM (Gygli et al., 2015)	0.415	0.467	0.469	0.478	0.445	0.458	0.451	0.407	0.473	0.453	0.461
SMRS (Elhamifar et al., 2012)	0.272	0.324	0.331	0.362	0.289	0.276	0.302	0.297	0.314	0.295	0.306
Quasi (Kim et al., 2014)	0.336	0.369	0.342	0.375	0.324	0.301	0.318	0.295	0.327	0.309	0.329
MBF (Chu et al., 2015)	0.295	0.357	0.325	0.412	0.318	0.334	0.365	0.313	0.365	0.357	0.345
CVS (Panda & Roy-Chowdhury, 2017)	0.328	0.413	0.379	0.398	0.354	0.381	0.365	0.326	0.402	0.378	0.372
SG (Mahasseni et al., 2017)	0.423	0.472	0.475	0.489	0.456	0.473	0.464	0.417	0.483	0.466	0.462
VESD (Cai et al., 2018)	0.447	0.493	0.496	0.503	0.478	0.485	0.487	0.441	0.492	0.488	0.481
DSN (Panda et al., 2017)	0.373	0.441	0.428	0.436	0.411	0.417	0.412	0.368	0.435	0.416	0.424
LM (Xiong et al., 2019)	0.559	0.429	0.612	0.540	0.604	0.475	0.432	0.663	0.691	0.626	0.563
MINI-Net (Hong et al., 2020)	0.785	0.566	0.736	0.753	0.769	0.633	0.612	0.756	0.756	0.656	0.702
LR (Ye et al., 2021)	0.850	0.714	0.819	0.786	0.802	0.755	0.716	0.773	0.786	0.681	0.768
CHD (Badamdorj et al., 2022)	-	-	-	-	-	-	-	-	-	-	0.528
RASL (Li et al., 2024)	0.897	0.657	0.925	0.583	0.847	0.830	0.697	0.833	0.875	0.890	<u>0.783</u>
Ours	0.824	0.741	0.872	0.788	0.806	0.725	0.758	0.837	0.880	0.787	**0.802**

Table 3: mAP on the YouTube Highlights dataset.

Method	Dog	Gymnastics	Parkour	Skating	Skiing	Surfing	Average
GIF (Gygli et al., 2016)	0.308	0.335	0.540	0.554	0.328	0.541	0.464
LSVM (Sun et al., 2014)	0.60	0.41	0.61	0.62	0.36	0.51	0.54
RARE (Yang et al., 2015)	0.49	0.35	0.50	0.25	0.22	0.49	0.38
LM (Xiong et al., 2019)	0.579	0.417	0.670	0.578	0.486	0.651	0.564
MINI-Net (Hong et al., 2020)	0.577	0.574	0.698	0.522	0.539	0.593	0.584
LR (Ye et al., 2021)	0.554	0.623	0.701	0.691	0.601	0.598	0.630
CHD (Badamdorj et al., 2022)	0.606	0.711	0.742	0.498	0.682	0.685	<u>0.654</u>
RASL (Li et al., 2024)	0.642	0.758	0.709	0.456	0.665	0.667	0.651
Ours	0.679	0.628	0.819	0.587	0.613	0.698	**0.670**

4.3 Comparison with state-of-the-art approaches

We compare our framework against a comprehensive range of classical and recent video highlight detection and summarization methods, spanning fully and weakly supervised paradigms. As shown

Table 4: mAP50 on the MoSu dataset. **T.** denotes Topic.

Method	T.1	T.2	T.3	T.4	T.5	T.6	T.7	T.8	T.9	T.10	Average
MINI-Net(Hong et al., 2020)	0.599	0.558	0.596	0.587	0.626	0.584	0.593	0.587	0.613	0.579	0.592
DL-VHD (Xu et al., 2021)	0.592	0.565	0.591	0.597	0.635	0.577	0.590	0.586	0.607	0.576	0.592
Ours	0.606	0.565	0.601	0.616	0.682	0.576	0.594	0.593	0.614	0.600	0.605

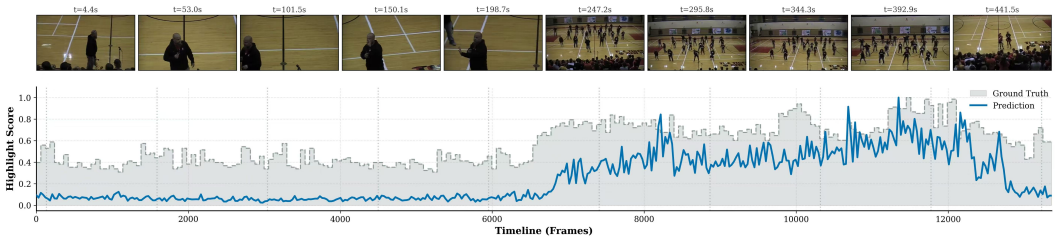

Figure 5: Highlight score predictions on the TVSum dataset for the flashmob topic. More results are in the appendix.

in Tab. 2, Tab. 3, and Tab. 4, our method establishes new state-of-the-art results across three datasets: the classic TVSum and YouTube Highlights, and the recently introduced, large-scale MoSu dataset.

On the classic benchmarks (Tab. 2 and Tab. 3), our approach achieves pioneering average mAP scores of $\mathbf{0.802}$ and $\mathbf{0.670}$, respectively, significantly outperforming existing methods by a substantial margin. Unlike earlier approaches that train small-scale models from scratch on isolated audio-visual features, our framework successfully unlocks the robust cognitive priors of MLLMs. By orchestrating a co-evolutionary loop between the generative director and the critic, we effectively overcome the limitations of stationary weakly-labeled datasets.

Furthermore, on the expansive MoSu dataset (Tab. 4), our method achieves a leading average mAP of $\mathbf{0.605}$. Notably, even when utilizing simple CLIP features provided, our approach consistently surpasses established baselines like MINI-Net. This performance gain on a diverse, large-scale dataset underscores the strong generalization and versatility of our proposed paradigm. Qualitative results (Fig. 5 and A.6.3) further demonstrate our model's capacity to generate semantically accurate highlight predictions using only video-level annotations.

5 Conclusion

In this study, we introduced a novel generative co-evolution post-training paradigm to adapt Multimodal Large Language Models (MLLMs) for weakly-supervised video highlight detection. By orchestrating a self-evolutionary process where a generative pipeline (an LLM Director and a text-to-video Film-maker) synthesizes adversarial video data to continuously challenge a discriminative Critic (the target MLLM), our framework achieves new state-of-the-art performance across multiple benchmarks with minimal human supervision.

Future research will focus on two key directions. First, we plan to extend this multi-agent training paradigm to other complex tasks within and beyond video understanding. Second, rather than relying on a frozen text-to-video model, we aim to simultaneously co-train the Film-maker alongside the Director and Critic. This dynamic optimization will allow for the generation of even more targeted and challenging visual samples, further boosting the MLLM for self-improvement.

6 AI Use Statement

During the preparation of this work, the authors used generative AI tools solely for the purpose of language polishing and grammar checking. After using these tools, the authors reviewed and edited the content as needed and take full ultimate responsibility for the content of this publication.

REFERENCES

Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Contrastive learning for unsupervised video highlight detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14022–14032. IEEE, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Sijia Cai, Wangmeng Zuo, Larry S Davis, and Lei Zhang. Weakly-supervised video summarization using variational encoder-decoder and web prior. In *European Conference on Computer Vision*, pp. 193–210. Springer, 2018.

Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3584–3592, 2015.

Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Russ R Salakhutdinov. Good semi-supervised learning that requires a bad gan. *Advances in neural information processing systems*, 30, 2017.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 1600–1607. IEEE, 2012.

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27, 2014.

Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 907–916. IEEE, 2026.

Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3090–3098, 2015.

Michael Gygli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1001–1009, 2016.

Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. In *European conference on computer vision*, pp. 345–360. Springer, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gunhee Kim, Leonid Sigal, and Eric P Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp. 4225–4232, 2014.

Sumin Kim, Hyemin Jeong, Mingu Kang, Yejin Kim, Yoori Oh, and Joonseok Lee. Triplesumm: Adaptive triple-modality fusion for video summarization. *arXiv preprint arXiv:2603.01169*, 2026.

Min Jung Lee, Dayoung Gong, and Minsu Cho. Video summarization with large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18981–18991. IEEE, 2025.

Tingtian Li, Zixun Sun, and Xinyu Xiao. Unsupervised modality-transferable video highlight detection with representation activation sequence learning. *IEEE Transactions on Image Processing*, 33:1911–1922, 2024.

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*, pp. 5971–5984, 2024.

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

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3042–3051, 2022.

Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2982–2991. IEEE, 2017.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23023–23033. IEEE, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Rameswar Panda and Amit K Roy-Chowdhury. Collaborative summarization of topic-related videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4274–4283. IEEE, 2017.

Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K Roy-Chowdhury. Weakly supervised summarization of web videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3677–3686. IEEE, 2017.

Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *European conference on computer vision*, pp. 540–555. Springer, 2014.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5179–5187, 2015.

Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*, pp. 787–802. Springer, 2014.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376, 2025.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 1195–1204, 2017.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Li, Kun Wan, and Wentian Zhao. Vision-zero: Scalable vlm self-evolution via multi-agent self-play. In *International Conference on Learning Representations*, volume 2026, pp. 106716–106744, 2026a.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE international conference on big data (bigdata)*, pp. 2247–2256. IEEE, 2023.

Tao Wu, Li Yang, Gen Zhan, Yabin Zhang, Yiting Liao, Junlin Li, Deliang Fu, Li Zhang, and Limin Wang. Tempr1: Improving temporal understanding of mllms via temporal-aware multi-task reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2756–2767, 2026.

Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18709–18719. IEEE, 2024.

Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1258–1267. IEEE, 2019.

Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. Cross-category video highlight detection via set-based learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7970–7979, October 2021.

Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proceedings of the IEEE international conference on computer vision*, pp. 4633–4641, 2015.

Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7930–7939. IEEE, 2021.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

Jintao Zhang, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E Gonzalez, Jianfei Chen, and Jun Zhu. Turbodiffusion: Accelerating video diffusion models by 100-200 times. *arXiv preprint arXiv:2512.16093*, 2025.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pp. 766–782. Springer, 2016.

Ping Zhang, Zheda Mai, Quang-Huy Percy Nguyen, and Wei-Lun Harry Chao. Revisiting semi-supervised learning in the era of foundation models. *Advances in Neural Information Processing Systems*, 38:59295–59324, 2026.

A APPENDIX

This appendix provides comprehensive implementation details for the core experimental settings evaluated in this study: (i) the preliminary supervised fine-tuning (SFT) setup, (ii) the Multiple Instance Learning (MIL) baseline, (iii) the semi-supervised learning (SSL) extension, and (iv) our proposed generative co-evolution paradigm. Additionally, we present extended qualitative results, visual examples of intermediate outputs, and a reproducibility statement.

The remainder of this appendix is organized as follows:

- **Section A.1:** Summary of Notation
- **Section A.2:** Preliminary Supervised Fine-Tuning Setup
  - **Section A.2.1:** Experimental Settings
  - **Section A.2.2:** Quantitative Analysis
- **Section A.3:** Multiple Instance Learning Baseline
- **Section A.4:** Semi-Supervised Learning Extensions
- **Section A.5:** Generative Co-Evolution Framework
  - **Section A.5.1:** Algorithmic Description
  - **Section A.5.2:** Implementation Details
- **Section A.6:** Extended Results
  - **Section A.6.1:** Per-Domain Results of the Ablation Study
  - **Section A.6.2:** Highlight Detection Outputs
  - **Section A.6.3:** Script Generation Examples
- **Section A.7:** Reproducibility Statement

Section Structure Summary. Section A.1 consolidates mathematical notations and abbreviations used throughout the manuscript. Section A.2 outlines the experimental setup for applying standard SFT to weakly-supervised video highlight detection (WSVHD). Section A.3 presents the architecture and adaptation details of our multimodal large language model (MLLM) baseline under the MIL framework. Section A.4 elaborates on the integration of unlabeled data via a Mean-Teacher framework. Section A.5 provides algorithmic formulations and training procedures for the generative co-evolution approach. Finally, Section A.6 showcases extended qualitative analysis, while Section A.7 details the compute environment, hyperparameters, and assets required to reproduce our findings.

A.1 SUMMARY OF NOTATION

Table 5 summarizes the notation used in the main text and implementation details. Following the main text, $G(\cdot)$ denotes the Film-maker, whereas G denotes the GRPO group size. The symbol α denotes the EMA momentum in the Mean Teacher baseline and the highlight-reward weight in co-evolution. The SFT frame count and LoRA scaling parameter are denoted by K_{frm} and α_{LoRA}, respectively.

Table 5: Symbols and their meanings.

Symbol	Meaning
Data and representations	
$\mathbb{D}, \mathbb{D}_s$	Weakly labeled real-video dataset and synthetic-video dataset, respectively.
N, K	Numbers of real video bags in $\mathbb{D}$ and synthetic clips in $\mathbb{D}_s$, respectively.
$\mathbb{V}_i, \mathbb{V}_k^s$	Real video i and synthetic 1-second video clip k, respectively.
y_i	Binary video-level supervision label, with 1 indicating highlight presence and 0 indicating absence.

Symbol	Meaning
i, j, k, t	Video indices (i, j), synthetic clip index (k), and temporal instance index (t).
T_i	Number of temporal instances in real video i.
d_v, d_a, d	Visual-encoder, audio-encoder, and shared representation dimensions, respectively.
$\boldsymbol{V}_i, \boldsymbol{v}_{i,t}$	Input instance-feature matrix and feature vector of its t-th instance.
$\boldsymbol{V}_i^{(v)}, \boldsymbol{V}_i^{(a)}$	Visual and audio feature matrices, respectively.
$\boldsymbol{V}_i^{(v)\prime}$	Visual features after projection into the shared feature space.
$\boldsymbol{W}_v, \boldsymbol{b}_v$	Weight matrix and bias of the visual feature projection.
$\tilde{\boldsymbol{V}}_i^{(v)}, \tilde{\boldsymbol{V}}_i^{(a)}$	Visual and audio representations after modality-specific self-attention.
$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$	Query, key, and value matrices in cross-attention.
$\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V$	Query, key, and value projection matrices.
$\boldsymbol{Z}_i, \boldsymbol{z}_{i,t}$	Cross-modally fused instance-feature matrix and its t-th instance representation.
MIL baseline	
$h, \boldsymbol{s}_i$	Highlight scoring function and the resulting temporal score vector.
$s_{i,t}$	Highlight score of the t-th instance in video i.
$\boldsymbol{z}_i^{\text{bag}}$	Video-level representation obtained by pooling instance features with highlight-score weights.
$\hat{y}_i, \sigma(\cdot)$	Predicted bag-level highlight probability and sigmoid function, respectively.
$\mathbb{B}$	Training mini-batch.
$\mathbb{B}^+, \mathbb{B}^-$	Positive- and negative-video index sets within a mini-batch.
$\mathcal{L}_{\text{CE}}, \mathcal{L}_{\text{rank}}$	Binary video-level classification loss and pairwise instance-ranking loss.
$\mathcal{L}_{\text{base}}$	MIL objective: $\mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{rank}}$.
Δ, λ	Ranking margin and ranking-loss weight, respectively.
Mean Teacher semi-supervised learning	
$\mathbb{B}_u$	Mini-batch of unlabeled real videos used in the SSL baseline.
$\boldsymbol{\theta}^S, \boldsymbol{\theta}^T$	Student and teacher MLLM parameter vectors, respectively.
$\boldsymbol{s}_i^S, \boldsymbol{s}_i^T$	Student and teacher instance-level highlight-score vectors.
$\hat{y}_i^S, \hat{y}_i^T$	Student and teacher bag-level highlight probabilities.
$\boldsymbol{p}_i^S, \boldsymbol{p}_i^T$	Student and teacher binary bag-level probability distributions.
$\mathcal{L}_{\text{MT}}$	Mean Teacher consistency loss combining instance-level MSE and bag-level KL divergence.
β	Weight of the bag-level KL-divergence term in $\mathcal{L}_{\text{MT}}$.
α	EMA momentum used to update the teacher parameters.
Generative co-evolution	
$\pi_{\boldsymbol{\theta}}$	Director policy conditioned on the context prompt.
$\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}}$	Current and previous Director parameter vectors.
p, z, z_g	Context prompt, generated scene script, and the g-th script in a sampled group.
$G(\cdot)$	Frozen Film-maker that renders a scene script into a 1-second synthetic video clip.
G, g	GRPO group size and candidate index within a group.
$\mathbb{V}_g^s, \mathbb{V}_k^s$	Synthetic 1-second video clips for candidate g and dataset entry k, respectively.
c	Critic class index: 0 for non-highlight, 1 for highlight, and 2 for fake.
c_g, c_k	Binary generation condition labels (1 for highlight, 0 for non-highlight).
$\hat{\boldsymbol{y}}, \hat{y}^{(c)}$	Critic three-class probability vector and its component for class c.
s_g, s_k	Instance-level highlight scores predicted by the Critic for clips g and k.
r_g^{tf}	Bag-level true–fake reward: $r_g^{tf} = 1 - \hat{y}_g^{(2)}$.
r_g^{hl}	Condition-aligned highlight reward: $r_g^{hl} = c_g s_g + (1 - c_g)(1 - s_g)$.
r_g, α	Composite reward $r_g = r_g^{tf} + \alpha r_g^{hl}$ and its highlight-reward weight.
μ_r, σ_r	Mean and standard deviation of rewards within a sampled group.
A_g	Group-normalized relative advantage: $A_g = (r_g - \mu_r)/\sigma_r$.

Symbol	Meaning
$\epsilon, \mathcal{J}_D$	GRPO clipping parameter and Director optimization objective.
$\mathcal{L}_{\text{sup}}$	Supervised highlight-classification loss on weakly labeled real videos.
$\mathcal{L}_{\text{fake}}$	Synthetic-video classification loss with fake-class target 2.
$\mathcal{L}_{\text{ins}}$	Synthetic instance-level classification loss supervised by condition label c_k.
λ_f, λ_i	Weights of $\mathcal{L}_{\text{fake}}$ and $\mathcal{L}_{\text{ins}}$, respectively.
$\mathcal{L}_C$	Critic objective: $\mathcal{L}_{\text{sup}} + \lambda_f \mathcal{L}_{\text{fake}} + \lambda_i \mathcal{L}_{\text{ins}}$.
$\mathcal{B}_{\text{sup}}, \mathcal{B}_{\text{fake}}$	Real weakly labeled and synthetic mini-batches in the co-evolution algorithm.
R, r	Total number of co-evolution rounds and current round index.
SFT baselines	
K_{frm}	Number of sampled frames per SFT input; set to 16.
$r_{\text{LoRA}}, \alpha_{\text{LoRA}}$	LoRA rank and scaling parameter; set to 16 and 32, respectively.

A.2 Initial Supervised Fine-Tuning Attempt

This section details the preliminary baseline experiments applying off-the-shelf Supervised Fine-Tuning (SFT) routines to adapt Multimodal Large Language Models (MLLMs) for Weakly-Supervised Video Highlight Detection (WSVHD). To systematically evaluate performance across modalities, we benchmark both visual-only and audio-visual MLLM variants on the TVSum and YouTube Highlights datasets.

A.2.1 Model Architectures and Experimental Configurations

Model Configurations. For visual-only experiments (**V**), we employ Qwen3-VL-2B-Instruct. For audio-visual experiments (**V+A**), we adopt Qwen2.5-Omni-3B. Both architectures are fine-tuned via Low-Rank Adaptation (LoRA) using the `ms-swift` (v4.0.4) framework. Domain adaptation is conducted independently: a separate LoRA adapter is trained per topic within each dataset.

We formulate the WSVHD task as a Video Question Answering (VQA) problem, training the model to infer segment-level highlight confidence from video-level supervision. Videos belonging to a target domain are assigned the label `1`, while out-of-domain videos serve as negative examples with label `0`. All models share consistent data splits and preprocessing pipelines across experiments.

Architecture Adaptation Details.

- **Qwen3-VL-2B-Instruct:** The vision encoder and multimodal aligner remain frozen, with LoRA trainable parameters applied exclusively to the linear projection layers of the language backbone. For training, $K_{\text{frm}} = 16$ frames are uniformly sampled across the full video length. During inference, 16 frames are uniformly sampled within each evaluation window.
- **Qwen2.5-Omni-3B:** Adaptation is limited to the core language model component (the "thinker"). The visual encoder, audio encoder, and modality alignment interfaces are held frozen, while speech generation features are disabled. Inputs include both uniform video frame sequences (16 frames) and full-length raw audio files decoded in WAV format. Segment-level evaluation at inference extracts audio matching the target time boundaries.

Prompt Formulation. Training and inference prompt protocols are detailed in Appendix Prompts A.2.1–A.2.1. Domain placeholders (`{domain}`) are dynamically instantiated using the topic's textual query (e.g., *"changing a vehicle tire"* or *"surfing"*).

VISUAL-ONLY MODEL SFT: TRAINING

Input. Sixteen temporally ordered frames sampled uniformly from a training video.

User Prompt.

Do these temporally ordered video frames contain a highlight related to "{domain}"? Answer 1 for yes or 0 for no.

Target Answer. 1 for target-domain videos; 0 for out-of-domain videos.

VISUAL-ONLY MODEL SFT: INFERENCE

Input. Sixteen temporally ordered frames sampled uniformly within an evaluation segment.

User Prompt.
Estimate the probability that this video segment contains a highlight related to "{domain}", based on these temporally ordered frames. Output only a number between 0.00 and 1.00, with two decimal places.

Output Handling. Parsed as a floating-point highlight score within $[0.00, 1.00]$. Discrete endpoints (0 and 1) are accepted and cast to floating-point values.

VISUAL-AUDIO MODEL SFT: TRAINING

Input. Sixteen temporally ordered frames sampled uniformly from a complete training video, alongside its full-length audio stream.

User Prompt.
Do these temporally ordered video frames and the accompanying audio contain a highlight related to "{domain}"? Answer 1 for yes or 0 for no.

Target Answer. 1 for target-domain videos; 0 for out-of-domain videos.

VISUAL-AUDIO MODEL SFT: INFERENCE

Input. Sixteen temporally ordered frames sampled uniformly within an evaluation segment, alongside the segment's synchronized audio stream.

User Prompt.
Estimate the probability that this video segment contains a highlight related to "{domain}", based on these temporally ordered frames and the accompanying audio. Output only a number between 0.00 and 1.00, with two decimal places.

Output Handling. Output parsing follows the identical rules described in the visual-only protocol.

Optimization and Hyperparameters. Table 6 outlines shared training hyperparameters, and Table 7 presents model-specific execution parameters. Both visual-only and multimodal architectures share identical effective batch sizes, random seeds, and optimization step schedules.

Inference and evaluation. During inference, we use deterministic decoding with a temperature of zero and a maximum of 16 generated tokens. For evaluation, we use the same data lists and evaluation metrics as in other experiments.

A.2.2 Quantitative Analysis

The quantitative performance of the SFT-based VQA formulation on the TVSum and YouTube Highlights datasets is summarized in Table 8 and Table 9, respectively. Despite pairing a strong multimodal backbone with standard fine-tuning strategies, post-training performance remains suboptimal across domains.

These empirical results suggest that directly framing WSVHD as a conventional video QA task is inherently limited, primarily due to the absence of fine-grained, segment-level supervision and explicit temporal structure modeling. These findings motivate our subsequent design choices to develop a more principled adaptation framework for MLLMs in weakly-supervised highlight detection.

Table 6: Shared data statistics and settings for both SFT baselines. Counts and training budgets apply separately to each baseline.

Setting	TVSum	YouTube
Frames per input	16	16
Maximum image area	65,536	65,536
LoRA rank / scaling	16 / 32	16 / 32
LoRA dropout	0.05	0.05
Optimizer	Fused AdamW	Fused AdamW
Learning rate	2×10^{-5}	2×10^{-5}
Adam betas	$(0.9, 0.95)$	$(0.9, 0.95)$
Adam epsilon	10^{-8}	10^{-8}
Weight decay	0.1	0.1
Maximum gradient norm	1.0	1.0
Schedule / warmup ratio	Cosine / 0.05	
Effective batch size	8	8
Logical epochs	20	10
Total training examples	1,600	4,680
Optimizer steps	200	585
Precision / attention	BF16 / SDPA	BF16 / SDPA
Gradient checkpointing	Enabled	Enabled
Sequence packing	Disabled	Disabled
Random seed / data seed	42 / 42	42 / 42
Temperature / maximum new tokens	0 / 16	0 / 16

Table 7: Model-specific SFT configurations. VL denotes Qwen3-VL-2B-Instruct with visual-only inputs; Omni denotes Qwen2.5-Omni-3B with audio-visual inputs. Sequence length refers to the maximum training input length.

	VL		Omni	
Setting	**TVSum**	**YouTube**	**TVSum**	**YouTube**
Input modalities	V	V	V+A	V+A
Maximum sequence length	4,096	4,096	32,768	40,960
Micro-batch size	2	2	1	1
Gradient accumulation	4	4	8	8
Inference batch size	4	16	1	1

Table 8: Per-domain Top-5 mAP on the TVSum dataset. VL denotes the model with visual-only inputs; Omni denotes the model with visual-audio inputs. The better result in each row is boldfaced.

Topic	Domain	VL	Omni
BK	Beekeeping	0.7947	**0.8802**
BT	Attempting a bike trick	**0.6604**	0.5383
DS	Dog shows	**0.7944**	0.6688
FM	Flash mob gathering	**0.6601**	0.6327
GA	Grooming an animal	**0.8655**	0.8581
MS	Making a sandwich	**0.6161**	0.3670
PK	Parkour	0.4844	**0.7853**
PR	Parade	**0.7433**	0.6984
VT	Changing a vehicle tire	**0.8068**	0.7981
VU	Getting a vehicle unstuck	**0.8627**	0.7032
Average		**0.7288**	0.6930

Table 9: Per-domain mAP on the YouTube Highlights dataset. VL denotes the model with visual-only inputs; Omni denotes the model with visual-audio inputs. The better result in each row is boldfaced.

Domain	VL	Omni
Dog	**0.4830**	0.4276
Gymnastics	0.5634	**0.6158**
Parkour	0.6714	**0.7466**
Skating	**0.5411**	0.4012
Skiing	0.6510	**0.6644**
Surfing	0.5633	**0.6144**
Average	**0.5789**	0.5783

A.3 THE MIL BASELINE

This section describes the model and training details of our proposed MLLM adaption approach using MIL as shown in Figure 2 in the main text. The model backbone extracts frozen multimodal representations from pre-trained foundation models, which are subsequently processed by lightweight, learnable contextualization and scoring modules.

Feature Extraction Encoders. Visual representations are extracted at 1 frame-per-second via the vision encoder of Qwen-3-VL-2B (Bai et al., 2025), producing segment-level feature vectors of dimension $d_v = 2048$. Audio representations are derived from the audio encoder of Qwen-3-ASR-0.6B (Shi et al., 2026), yielding corresponding segment-level vectors of dimension $d_a = 1024$.

Modality Projection and Self-Attention. Visual features $\boldsymbol{V}_i^{(v)} \in \mathbb{R}^{2048\times T_i}$ are mapped to a 1024-dimensional space via linear projection: $\boldsymbol{V}_i^{(v)\prime} = \boldsymbol{W}_v\boldsymbol{V}_i^{(v)} + \boldsymbol{b}_v\mathbf{1}_{T_i}^\top$, where $\boldsymbol{W}_v \in \mathbb{R}^{1024\times 2048}$, $\boldsymbol{b}_v \in \mathbb{R}^{1024}$, and $\mathbf{1}_{T_i} \in \mathbb{R}^{T_i}$ is a vector of ones. To model intra-modal temporal dependencies, the visual and audio feature matrices are independently processed by single-layer Multi-Head Self-Attention (MHSA) blocks. Each block is followed by a Dropout layer, a residual connection, and Layer Normalization:

$$\tilde{\boldsymbol{V}}_i^{(v)} = \text{LayerNorm}\left(\boldsymbol{V}_i^{(v)\prime} + \text{Dropout}\left(\text{MHSA}_v(\boldsymbol{V}_i^{(v)\prime})\right)\right) \tag{4}$$

$$\tilde{\boldsymbol{V}}_i^{(a)} = \text{LayerNorm}\left(\boldsymbol{V}_i^{(a)} + \text{Dropout}\left(\text{MHSA}_a(\boldsymbol{V}_i^{(a)})\right)\right) \tag{5}$$

Both MHSA modules operate with a hidden dimension $d = 1024$, 8 attention heads, and a dropout probability of 0.1.

Cross-Modal Fusion Block. Cross-modal interactions are established using a Multi-Head Cross-Attention (MHCA) block. Visual features act as queries $\boldsymbol{Q} = \boldsymbol{W}_Q\tilde{\boldsymbol{V}}_i^{(v)}$, while audio features provide the keys $\boldsymbol{K} = \boldsymbol{W}_K\tilde{\boldsymbol{V}}_i^{(a)}$ and values $\boldsymbol{V} = \boldsymbol{W}_V\tilde{\boldsymbol{V}}_i^{(a)}$, where $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{1024\times 1024}$. The fused, contextualized representation matrix $\boldsymbol{Z}_i = [\boldsymbol{z}_{i,1}, \boldsymbol{z}_{i,2}, \ldots, \boldsymbol{z}_{i,T_i}] \in \mathbb{R}^{1024\times T_i}$ is computed as:

$$\boldsymbol{Z}_i = \text{LayerNorm}\left(\tilde{\boldsymbol{V}}_i^{(v)} + \text{Dropout}\left(\text{MHCA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})\right)\right) \tag{6}$$

The MHCA block similarly utilizes 8 attention heads and a 0.1 dropout rate.

Prediction Heads. The contextualized segment vectors $\boldsymbol{z}_{i,t} \in \mathbb{R}^{1024}$ are routed to two task-specific prediction modules:

1. *Instance-Level Score Head (MLP_{ins}):* A two-layer MLP estimating the segment highlight score $s_{i,t} \in [0, 1]$. It consists of a linear projection from 1024 to 64 dimensions, Layer Normalization, a Dropout layer ($p = 0.5$), a final linear projection to a scalar, and a Sigmoid activation $\sigma(\cdot)$.
2. *Bag-Level Classification Head (MLP_{bag} / MLP_{vid}):* For the MIL baseline and semi-supervised Mean Teacher configurations, MLP_{bag} applies a single linear projection from the 1024-dimensional aggregated bag representation to 2 output logits, normalized by a

Softmax activation. In the generative co-evolution setting, MLP_{vid} projects to 3 output logits corresponding to the expanded class space (0: non-highlight, 1: highlight, 2: synthetic/fake).

Training hyper-parameters. The MIL baseline is implemented in by adapting the MMAction training framework [2]. At each iteration, the number of sampled instances for each training video (the bag size) is 60 following Hong et al. (2020). The optimizer is Stochastic Gradient Descent (SGD) with a momentum of 0.9. The initial learning rate of 0.005 is decayed by a factor of 0.5 at epochs 100, 150, 200, and 250. Training runs for 300 epochs with a batch size of 128, weight decay of 5×10^{-5}, and gradient clipping with a maximum norm of 40. The ranking margin is 1.0 and the ranking-loss weight is 1.0. These settings are summarized in Table 10.

A.4 Semi-Supervised Learning Integration

This section details the empirical setup for integrating the Mean-Teacher framework into the MIL baseline, as presented in Section 3.3.

For each target topic, we incorporate 25 curated unlabeled videos to compute the semi-supervised regularization losses. Figure 6 illustrates random frame samples from the curated unlabeled subset across each topic in the TVSum dataset. The teacher model shares an identical architecture with the student model and is updated via an exponential moving average (EMA) with a momentum parameter of 0.9. The bag-level consistency weight β is set to 1.0. All remaining optimization parameters match the configuration of the MIL baseline.



Figure 6: Sample frames from the unlabeled video dataset utilized in the SSL experiments. Each row depicts representative frames from a single video corresponding to topics VT, VU, GA, MS, PK, PR, FM, BK, BT, and DS, respectively.

[2] https://github.com/open-mmlab/mmaction2

A.5 Generative Co-Evolution Framework

This section provides the experimental configurations for the generative co-evolution paradigm introduced in Section 3.4. The structural specification of the Critic module aligns directly with the MIL baseline.

A.5.1 Co-Evolution Algorithm

Algorithm 1 formalizes the alternating optimization procedure executed between the Director and the Critic. During this phase, the Critic omits unsupervised real samples to evaluate exclusively weakly-labeled real bags in conjunction with dynamically generated synthetic clips.

Algorithm 1 Generative Co-Evolution Framework for WSVHD

Require: Weakly-labeled real dataset $\mathbb{D} = \{(\mathbb{V}_i, y_i)\}_{i=1}^{N}$; Director $\pi_{\boldsymbol{\theta}}$; Film-maker G; Critic MLLM.
Require: Co-evolution rounds R, GRPO group size G, synthetic clips per round K, reward weight α, loss weights λ_f, λ_i.
for $r = 1$ **to** R **do**
  Phase 1: Director Update (GRPO)
  Sample context prompt p and clip condition labels $\{c_g\}_{g=1}^{G} \subset \{0,1\}^G$.
  Generate a group of candidate scripts $\{z_g\}_{g=1}^{G} \sim \pi_{\boldsymbol{\theta}}(\cdot \mid p)$.
  for $g = 1$ **to** G **do**
    Render 1-second synthetic video clip $\mathbb{V}_g^s \leftarrow G(z_g)$.
    Obtain Critic predictions for the group bag: class probabilities $\hat{\boldsymbol{y}}_g \in [0,1]^3$ and instance score s_g.
    Compute bag-level realism reward: $r_g^{tf} \leftarrow 1 - \hat{y}_g^{(2)}$.
    Compute instance-level highlight reward: $r_g^{hl} \leftarrow c_g s_g + (1 - c_g)(1 - s_g)$.
    Compute total composite reward: $r_g \leftarrow r_g^{tf} + \alpha r_g^{hl}$.
  end for
  Compute relative advantages $A_g \leftarrow (r_g - \mu_r)/\sigma_r$ across group $\{r_g\}_{g=1}^{G}$.
  Update Director $\pi_{\boldsymbol{\theta}}$ by maximizing the GRPO surrogate objective $\mathcal{J}_D(\boldsymbol{\theta})$.

  Phase 2: Synthetic Data Generation
  Synthesize K targeted clips using the updated Director $\pi_{\boldsymbol{\theta}}$ and Film-maker G:
  $\mathbb{D}_s \leftarrow \{(\mathbb{V}_k^s, c_k)\}_{k=1}^{K}$, where $\mathbb{V}_k^s = G(z_k)$, $z_k \sim \pi_{\boldsymbol{\theta}}(\cdot \mid p_k)$, and $c_k \in \{0,1\}$.

  Phase 3: Critic Update
  Sample mini-batches from real dataset $\mathbb{D}$ and synthetic dataset $\mathbb{D}_s$.
  Compute composite Critic loss: $\mathcal{L}_C \leftarrow \mathcal{L}_{\text{sup}} + \lambda_f \mathcal{L}_{\text{fake}} + \lambda_i \mathcal{L}_{\text{ins}}$.
  Update Critic parameters via gradient descent to minimize $\mathcal{L}_C$.
end for
return Optimized Target Critic MLLM.

A.5.2 Implementation Details

Director and Filmmaker Components. In the generative co-evolution phase, the Director is implemented using the Qwen3-0.6B language model and fine-tuned via Group Relative Policy Optimization (GRPO) using the HuggingFace TRL library.[3] The Filmmaker module utilizes the Wan2.1-1.3B text-to-video generation model, accelerated by the TurboDiffusion framework[4] to synthesize 1-second video clips (16 frames per clip) at a spatial resolution of 480p.

Optimization and Generation Settings. The GRPO training configuration employs a maximum completion length of 1024, 16 generations per group, and 100 optimization steps per phase. Video synthesis via the Filmmaker operates with `sla_topk` $= 0.8$ across 4 sampling steps. During training, the Critic processes both weakly-labeled real video bags and newly synthesized clips. To maintain balanced optimization, one synthetic bag is sampled per real training bag, matching the batch sizes between real and generated streams. Each synthetic bag contains between 10 and 50 generated clips, with a fixed 20% ratio of positive highlight clips. Table 10 summarizes the hyperparameter configurations across all MIL-based training paradigms.

[3] https://github.com/huggingface/trl
[4] https://github.com/thu-ml/TurboDiffusion

Table 10: Hyperparameter configurations for the MIL baseline, MeanTeacher experiments, and the generative co-evolution framework.

Hyperparameter	Value
Critic / MIL Baseline Optimization	
Learning Rate (lr)	0.005
SGD Momentum	0.9
Weight Decay	5×10^{-5}
Gradient Clipping (max_norm)	40
Total Epochs	300
Batch Size $\lvert\mathbb{B}\rvert$	128
Bag Size	60
Ranking Margin Δ	1.0
Ranking Loss Weight λ	1.0
Mean Teacher SSL	
Unlabeled Videos per Topic	25
Exponential Moving Average (α)	0.9
Consistency Loss Weight β	1.0
Director Optimization (GRPO)	
Max Completion Length	1024
Group Size (G)	16
Steps per Evolution Phase	100
GRPO Clipping Parameter (ϵ)	0.2
Instance Reward Weight (α)	0.5
Film-maker Generation	
SLA Top-k Ratio	0.8
Sampling Steps	4
Generated Frames per Clip	16

Director Prompts. The Director ($\pi_{\boldsymbol{\theta}}$) autoregressively generates structured textual scripts conditioned on prompt instructions. These scripts serve as prompt inputs for the Filmmaker generator $G(z)$ during video synthesis. For each topic, two distinct prompt templates are employed: a *positive prompt* to synthesize scripts depicting highlight moments, and a *negative prompt* to synthesize scripts describing out-of-domain, non-highlight content.

As illustrated in Figure 7, the negative prompts share a unified structure across all categories, differing only in the target domain specification.

Director Negative Prompt

Generate a description for a video clip. This description will be used as prompt for a text-to-video model. The requirements are as follows:

- The video should not contain any content related to the domain of {`TOPIC`}.
- The generated video based on this description should be realistic.
- The description should include the appearance of major objects/humans/animals in the clip, as well as the actions they are performing.
- The description should include the description of the background environment.
- The description should be specific and concise. Assume the text-to-video model is not strong, and use commonly used language to describe the clip.

Figure 7: The negative prompt template provided as input to the Director LLM (Qwen3-0.6B) during generative co-evolution.

Similarly, positive prompts follow a coherent structural template across topics. The Director is first assigned a role and instructed to incorporate domain-typical subjects, actions, and environmental conditions. It is then constrained to adhere to a standardized descriptive style and paragraph structure, supported by in-context examples. The template layout for positive prompt generation is structured as follows:

DIRECTOR: SCRIPT GENERATION

COMMON ELEMENTS
{*Topic-specific subjects, objects, activities, environments, and visual characteristics.*}

SHOT & CAMERA
{*Typical shot types, viewpoints, framing, and camera movements appropriate to the topic.*}

LIGHTING & SETTING
{*Typical lighting conditions, locations, backgrounds, and relevant environmental details.*}

SUBJECT MATTER / SCENARIO TYPES (you can pick one from the examples below or propose yourself)
{*Topic-specific scenarios, including the main subjects, actions or interactions, relevant objects, settings, and descriptive keywords.*}

DESCRIPTIVE STYLE
- Use topic-appropriate vocabulary: {*DESCRIPTIVE_WORDS_AND_ACTION_VERBS*}.
- Describe the main subjects, their actions, and relevant visual details clearly and concretely.
- Follow these topic-specific conventions: {*DETAIL_AND_STYLE_GUIDELINES*}.

STRUCTURE OF EACH DESCRIPTION (4–6 sentences)
1. Establish the main subject and setting.
2. Describe the core action, movement, or interaction.
3. Add relevant details about appearance, objects, or the progression of the action.
4. Describe the surrounding environment, lighting, and camera behavior when relevant.
5. Conclude with the outcome, overall atmosphere, or what the video captures.

TONE
- Use objective, observational language without unnecessary emotional exaggeration.
- Follow the topic-appropriate tone: {*TONE_GUIDELINES*}.
- Write in English. Each description should be a single paragraph of 4–6 sentences.

Some examples of video description are shown below:

1. {*EXAMPLE_1*}
2. {*EXAMPLE_2*}
3. {*EXAMPLE_3*}
4. {*EXAMPLE_4*}
5. {*EXAMPLE_5*}

Generate one video description now.

The complete textual prompt files provided to the Director are included as plain-text assets in the supplementary material.

A.6 Extended Qualitative and Quantitative Results

This section provides additional qualitative evaluations and visual demonstrations of the proposed framework.

A.6.1 Per-Domain Results of the Ablation Study

Table 11 provides the detailed per-domain quantitative evaluation corresponding to the overall ablation study in Table 1, covering 70 individual experimental evaluations across all ten domains.

A.6.2 Highlight Prediction Curves

Figures 8, 9, and 10 present visual comparisons between ground-truth annotations and our predicted segment-level highlight confidence curves on the TVSum and YouTube Highlights datasets. These visualizations demonstrate that the model accurately localizes highlight segments even in the absence of frame-level supervision during training.

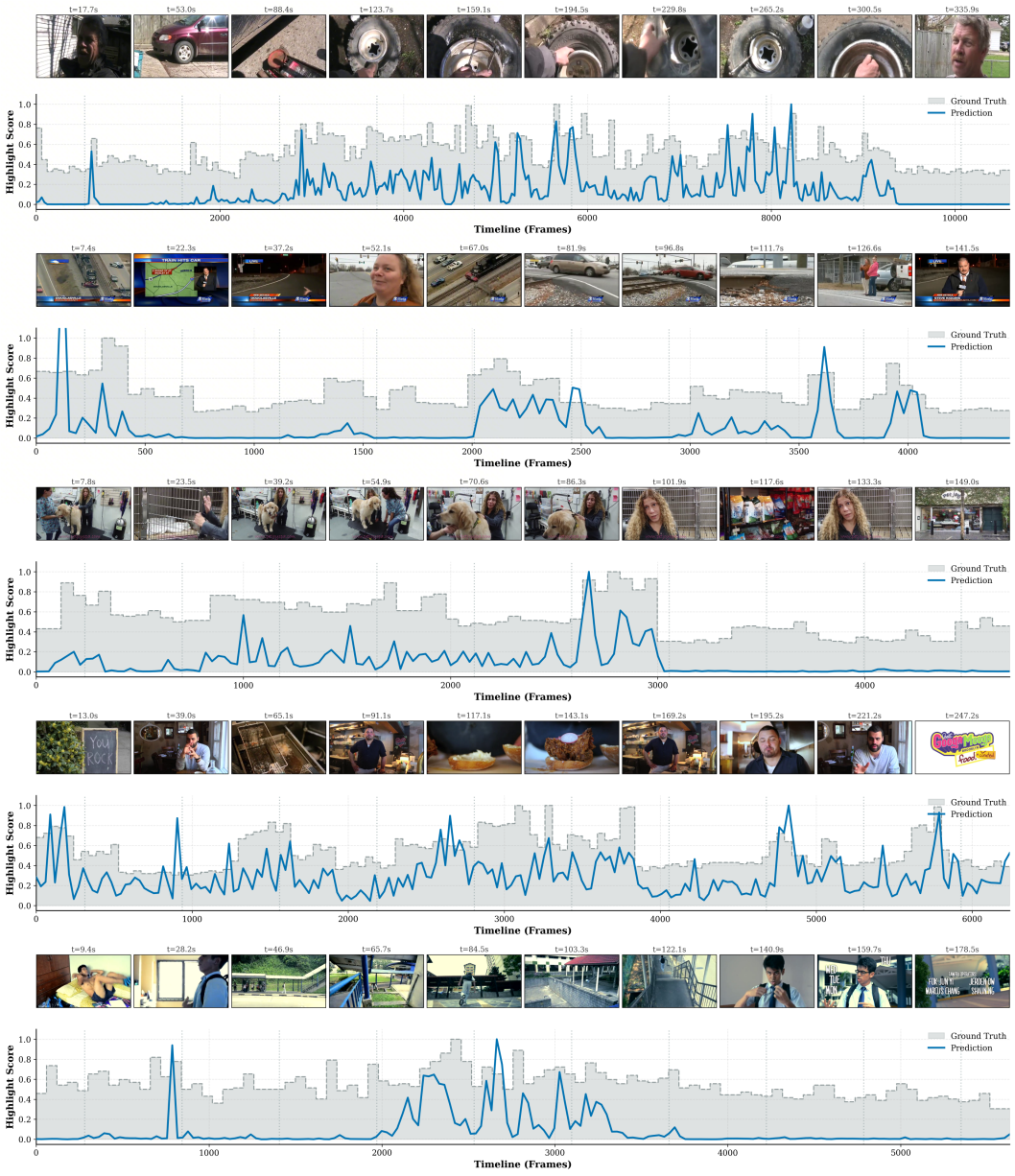

Figure 8: Qualitative highlight prediction curves on the TVSum dataset for topics VT, VU, GA, MS, and PK.



Figure 9: Qualitative highlight prediction curves on the TVSum dataset for topics PR, FM, BK, BT, and DS.

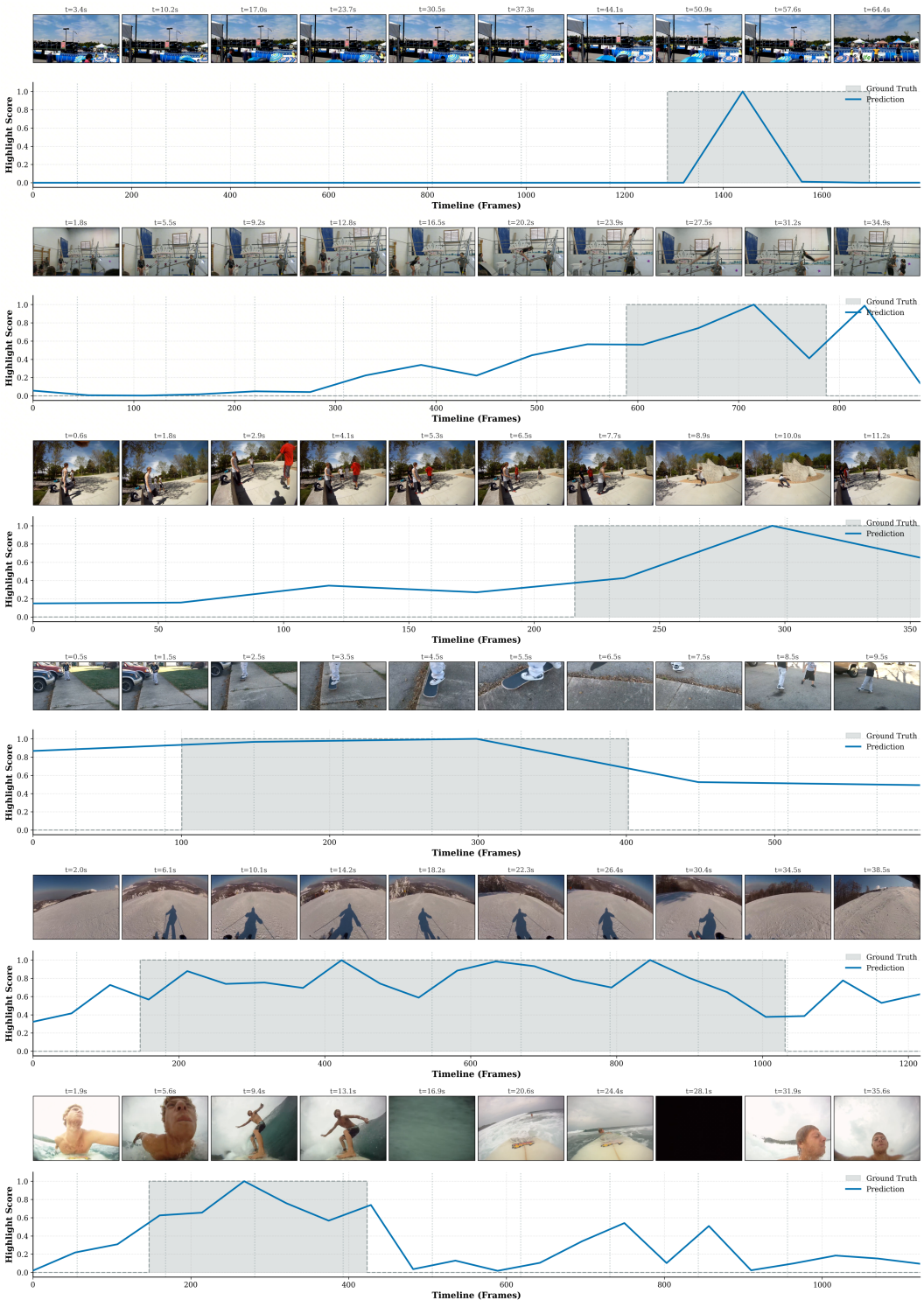

Figure 10: Qualitative highlight prediction curves on the YouTube Highlights dataset across all six categories: dog, gymnastics, parkour, skating, skiing, and surfing.

Table 11: Experimental results on the TVSum dataset (Top 5 mAP). Best and second-best are **bold-faced** and underlined, respectively.

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Average
MLLM + MIL	0.804	0.743	0.763	0.779	0.778	0.751	0.726	0.796	0.785	0.757	0.768
MLLM + MIL + SSL	0.832	0.720	0.784	0.799	0.822	0.744	0.738	0.817	0.754	0.723	0.773
Ours (no director update)	0.830	0.713	0.845	0.774	0.783	0.701	0.772	0.781	0.845	0.738	0.778
Ours (Bag-only Reward)	0.822	0.740	0.818	0.789	0.775	0.785	0.742	0.800	0.807	0.748	0.782
Ours (Full, 100 clips)	0.837	0.761	0.854	0.792	0.784	0.790	0.756	0.844	0.791	0.792	0.800
Ours (Full, 300 clips)	0.824	0.741	0.872	0.788	0.806	0.725	0.758	0.837	0.880	0.787	**0.802**
Ours (Full, 500 clips)	0.835	0.742	0.862	0.812	0.779	0.723	0.734	0.785	0.817	0.797	0.788

A.6.3 Synthesized Scripts and Generated Video Samples

Table 12 provides representative examples of positive and negative scripts synthesized by the Director across multiple topics. The Director consistently crafts descriptions of topic-relevant highlight moments for positive prompts, while accurately generating domain-irrelevant background content for negative prompts.

Figures 11 and 12 display randomly sampled synthetic video clips generated during training on the TVSum and MoSu datasets, respectively. Conditioned on the synthesized scripts, the video generator produces photorealistic and semantically aligned clips. These generated samples serve as effective positive and hard-negative data points, thereby improving the Critic's discriminative boundary.



Figure 11: Randomly sampled synthetic video clips generated during training on the TVSum dataset. Each row corresponds to representative clips synthesized for a target topic.

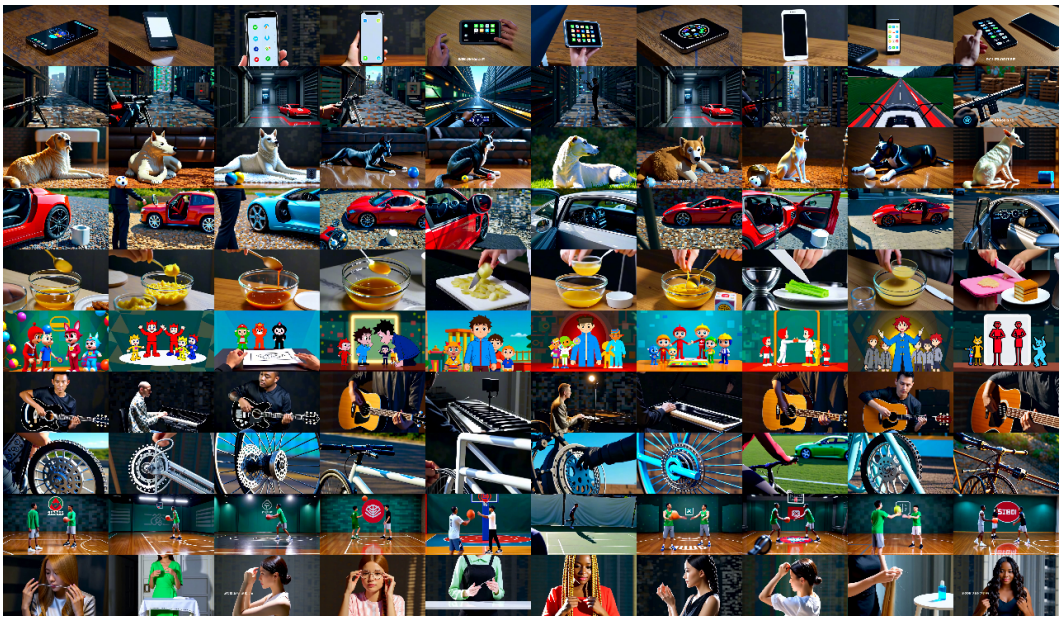

Figure 12: Randomly sampled synthetic video clips generated during training on the MoSu dataset. Each row corresponds to representative clips synthesized for a target topic.

Table 12: Examples of synthetic scene scripts generated by the Director model.

Domain	Condition	Generated Synthetic Script (z)
BK	Highlight	*"A close-up shot of a bustling beekeeping activity, with a slow-motion camera capturing the dynamic movement of bees around a hexagonal honeycomb. The bees, with golden-yellow black stripped bodies and dark brown fuzzy bodies, are seen densely packed, actively crawling over the cells filled with golden honey and pollen. Their movement is evident as they cluster around the hive, flying in intricate patterns, while some are collecting nectar. The camera maintains a steady, low-angle shot to emphasize the bees' coordinated efforts. The video is shot in natural daylight, highlighting the vibrant colors of the bees and the hive's narual beauty. The setting includes a green grassy field with a wooden fence, while the lighting creates a warm, inviting atmosphere. The video captures the industrious nature of the bees and the harmony of the environment."*
	Non-Highlight	*"A serene beach scene with a group of children playing in the sand, surrounding by palm trees and gentle waves. A young girl holds a small toy boat, while adults are fishing nets on the shore. The background is a sunny day with soft lighting, a distant ocean, and a shaded palm grove. The environment is calm and filled with natural beauty."*
BT	Highlight	*"A person in plaid shirt or camouflage patterned attire, wearing dark shorts, performs a bicycle wheelie on a wheelie cycle in a tall, sparsely lit warehouse. The camera follows the rider, capturing their balance as the wheel rises to the ceiling while the rider leans forward, maintaining steadiness with each movement. The background features high ceilings, corrugated metal walls, and open storefronts with cargo stacks, highlighting the warehouse's industrial energy. The wheelie lands smoothly, showing the rider's precision as they steady themselves, while a crowd of onlookers watch from behind metal barriers. The video conveys the thrill of motion and skill, emphasizing athleticism and dynamic energy in a cozy, modern setting."*

Table 12: Examples of synthetic scene scripts generated by the Director model (continued).

Domain	Condition	Generated Synthetic Script (z)
	Non-Highlight	*"A sunny afternoon in a quiet park where a young child, wearing a bright blue sports jacket and a shiny red bicycle, swings with it. The scene features a green tree, a yellow grass field, and a small spring for play. The child hops along the path, riding the bike with a smooth curve as they explore the space. The background adds soft light and a sense of movement, with the child laughing and pointing at things they find. The description is concise, focusing on the realistic interaction between the child, the environment, and the bike."*
DS	Highlight	*"A German Shepherd with a muscular, alert posture stands on a green platform, its black-and-tan coat gleaming under the sun. The handler, a man in orange tracksuit sunglasses, guides the dog with a confident stride, while the dog runs in low crouched position, following a blue cone. Behind it, a red wall and a white tent frame the scene, with a group of spectators seated near the platform. The dog's expressive eyes and wagging tail suggest a sense of purpose, while the background features a pink wall and a framed painting. Text overlays appear in bold, clean font, reading "DOG'S TALE" and "RELATO CANINO," emphasizing the dog's joyful energy. The overall atmosphere is calm and focused, capturing the dynamic interaction between the handler and the spirited dog."*
	Non-Highlight	*"A serene meadow is depicted with a white-tailed deer grazing. A young woman in a blue dress walks beside the deer, her fingers brushing against its fur. A small red bird watches from a nearby tree. The background is a lush green field with fluffy grass and a nearby stream, filled with natural beauty and quiet tranquility."*
MS	Highlight	*"A close-up shot of a freshly made sandwich, golden-brown on the bread and golden-brown on the sides, appears to be a simple assembly of ingredients. The person, wearing a white shirt with a red tie, carefully arranges a leaf of lettuce, a whole-wheat bun, and a cheese slice on the plate, their hands visible as they layer the sandwich. The camera slowly zooms in on the sandwich, highlighting its crisp texture and vibrant colors, while the kitchen counter is cluttered with tools and ingredients. The scene transitions to a casual atmosphere, capturing the methodical nature of the preparation with a warm, inviting tone."*
	Non-Highlight	*"A serene field in the morning, a vibrant green grass surrounded by a cluster of tulips and wildflowers. A young farmer, smiling, stands before a basket of ripe peaches and strawberries, hands guiding them with a small crutch. The scene includes a basket of colorful fruits and a group of children playing with pickling tools, their laughter blending with the music of a nearby tree. The background features a modest market square with stalls displaying fresh produce and a cheerful market stall selling sandwiches (not mentioned in the prompt). The description includes specific objects (fruits, tools), actions (planning, gathering), appearance details, and a realistic, concise setting."*

A.7 Environment and Reproducibility Statement

Computing Infrastructure. The MIL-based training and synthesis pipelines were conducted on a high-performance compute cluster equipped with $4\times$ NVIDIA A100 (80GB PCIe) GPUs and 512GB System RAM.

Software Environment. For the MIL-based experiments and video synthesis, code is implemented using Python 3.11.2, PyTorch v2.7.1, HuggingFace Transformers v4.57.6, and HuggingFace TRL (Transformer Reinforcement Learning) v0.29.1. Video generation acceleration for Wan2.1-1.3B is compiled using the official TurboDiffusion release with NVIDIA Driver version 535.261.03 and CUDA 12.6.

Reproducibility Statement. To support open scientific reproducibility, all source code, pre-extracted MLLM feature tensors for YouTube Highlights, TVSum, and MoSu, model checkpoints, and training script configurations will be publicly released upon publication.