From Script to Shot: A Benchmark for Grounding Screenplays in Movies

Jungu Cho[1,2], Young-Jae Park[3], Seong Jong Ha[1⋆], Siyeol Kim[4], Seungho Park[4], Jisu Shin[2], Junmyeong Lee[4], Chaemin Hwang[4], Hyunjun Jung[2], Sangeyl Lee[4], and Hae-Gon Jeon[4⋆]

[1] AI/DT Division, CJ Corporation
[2] Department of AI Convergence, GIST
[3] School of Information Convergence, Kwangwoon University
[4] Department of Artificial Intelligence, Yonsei University

Abstract. Aligning screenplay scenes to video shots is a foundational task for narrative video understanding. Unlike subtitles or captions, screenplays combine dialogue, visual direction, and narrative cues in a single document, making them ill-suited for standard video-text retrieval and leaving prior methods to exploit only subtitle overlaps. Methods ranging from alignment based on textual subtitle to contrastive encoders and temporal grounding networks can potentially bridge this gap, yet no benchmark systematically compares these paradigms or isolates the contributions of dialogue and visual information to alignment quality. We introduce "From Script to Shot", a benchmark of 50 feature films spanning nine decades with over 55K shot-level scene alignment annotations verified by human annotators. To disentangle text matching from visual grounding, we evaluate fourteen approaches—from sparse text retrieval to recent vision-language models such as FG-CLIP 2 and Qwen3-VL-Embedding, on an identical pipeline. Dialogue-based retrieval performs poorly on non-dialogue scenes, and contrastive encoders alone underperform direct text matching. Adaptive multimodal fusion achieves the strongest overall results, reaching 0.599 mSIoU when BM25 is fused with a Qwen3-VL-Embedding, yet even the strongest encoder leaves a consistent gap between dialogue and non-dialogue scenes, exposing the limits of current visual encoding for narrative grounding. By quantifying this gap, the benchmark provides a controlled testbed for studying screenplay-conditioned alignment. We demonstrate downstream utility through zero-shot movie scene segmentation and screenplay-guided video generation, confirming that the alignment signal generalizes beyond the benchmark itself. We release the full dataset, parser and parser outputs, toolkit, and baselines at `https://github.com/jungucho92/script2shot` to support research on long-form multimodal video understanding.

⋆ Corresponding authors

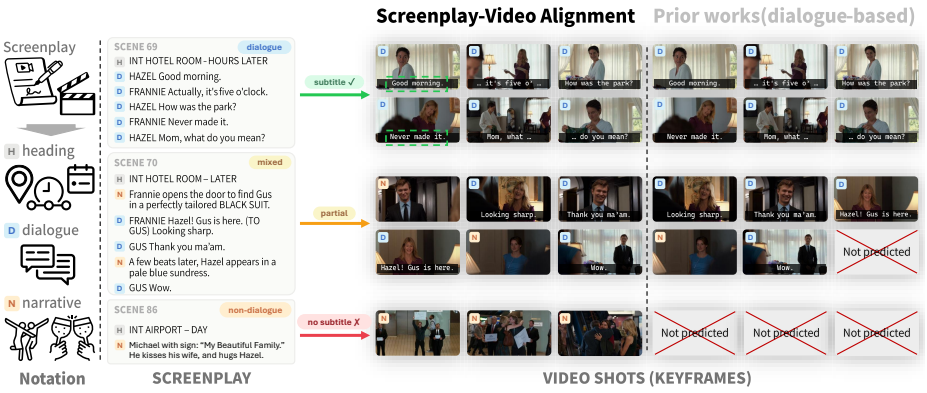

Fig. 1: Screenplay-video alignment maps each shot to its corresponding screenplay scene. Each scene combines a scene heading (H), narrative direction (N), and dialogue (D). Shots carry subtitle text only when spoken dialogue is present. Prior dialogue-based methods align subtitle-matched shots correctly but fail entirely on non-dialogue scenes, whereas our benchmark evaluates all scene types under a unified protocol.

1 Introduction

Understanding video in conjunction with language has progressed rapidly in recent years, yet most advances are driven by datasets [7, 23, 31, 45] composed of short, minute-scale clips paired with captions. While effective for localized action recognition [6, 36] or retrieval [3, 22], such datasets offer limited coverage of extended temporal dependencies and the higher-level narrative organization of a full story. Movies naturally exhibit this long-form structure, but their duration and complexity have limited the availability of large-scale, systematically annotated resources.

A movie is typically decomposed into scenes, narrative units defined by continuity of time, location, or action, each of which is realized on screen as a sequence of shots [17]. A screenplay describes the film as an ordered sequence of these scenes [4]. Each scene entry consists of a heading that sets the location and time, stage directions that describe actions or camera movement, and dialogue between characters. This mix of visual descriptions and spoken dialogue within a single document makes screenplays far richer than either subtitle transcripts or descriptive captions, each of which covers only one of the two channels [2]. Prior works [8, 14] on screenplay-video alignment have accordingly relied on matching subtitle text against dialogue lines, treating the task as a text overlap problem, yet the visual and narrative content that screenplays encode beyond dialogue remains unexploited.

Video-text retrieval aims to match video clips with natural-language descriptions, and vision-language models [22] trained for this task have achieved strong results on caption-level benchmarks [23,31,45]. However, screenplay text is structurally more complex than a caption: dialogue, stage directions, and narrative cues coexist within each scene, and it is unclear whether models designed for homogeneous captions can handle such heterogeneous input. Repurposing these models as baselines for screenplay-to-shot alignment demands a benchmark that

evaluates them under such conditions and separates the contributions of the dialogue and visual channels.

To address this gap, we introduce *From Script to Shot*, the first benchmark for screenplay-to-shot alignment in feature films, comprising 50 movies and over 55K human-verified shot-level annotations. The dataset captures the inherently ambiguous, one-to-many mapping between script scenes and cinematic shots, and is paired with a unified evaluation protocol that separates dialogue matching from visual grounding. We further provide standardized baselines across alignment methods and vision–language models, enabling controlled analysis of screenplay-conditioned alignment.

Through the proposed benchmark, we show that dialogue-based retrieval methods perform competitively overall but degrade substantially on non-dialogue scenes, consistent with their reliance on lexical matching, while vision–language encoders trained on caption-level data do not consistently surpass text-based alignment, suggesting limited transfer to heterogeneous screenplay inputs. Multimodal fusion yields the strongest aggregated performance, yet a persistent gap between dialogue-driven and visually grounded scenes remains even under the best configuration. These findings underscore the need for controlled evaluation that separates modality contributions and position the benchmark as a testbed for studying long-form screenplay-to-shot grounding.

We demonstrate that the alignment signal generalizes beyond the benchmark through two downstream applications. The induced scene boundaries serve as zero-shot pseudo-labels for scene segmentation without task-specific training, confirming that alignment captures structure invisible to visual-only detectors. Aligned scene–shot pairs enable an LLM to rewrite raw screenplay text into visually precise generation prompts, recovering film-specific cinematographic identity across three commercial text-to-video services.

Contributions.

- **A benchmark** for screenplay-to-shot alignment: 50 feature films spanning nine decades, with over 55K shot-level scene-alignment annotations on a controlled pipeline that isolates the dialogue and visual channels.
- **A systematic comparison** of fourteen approaches, from sparse text retrieval and contrastive encoders to recent vision-language embeddings (FG-CLIP 2, Qwen3-VL-Embedding), under identical decoding and metrics.
- **A quantitative diagnosis**: a persistent dialogue/non-dialogue grounding gap that survives even the strongest encoder (IoU 0.654 vs. 0.496), pinpointing visual encoding as the bottleneck for non-dialogue scenes.
- **Downstream evidence** that the alignment signal generalizes, via zero-shot scene segmentation and screenplay-guided video generation.
- **A full open release**: dataset, screenplay parser and its outputs, alignment toolkit, and baselines.

2 Related Works

2.1 Video-Text Retrieval

Video–Text Retrieval (VTR) aims to identify matching video–text pairs by bridging the semantic gap between video frames and natural language descrip-

tions through learning a shared embedding space. CLIP [27] marked a significant paradigm shift in VTR. Trained on 400 million image–text pairs, CLIP achieves highly effective visual–semantic alignment. CLIP4Clip [22] leverages CLIP's pre-trained image and text encoders for video–text retrieval. However, CLIP does not explicitly model sequential information in videos, limiting its ability to capture long-term and complex video narratives. DiffusionRet [16] formulates retrieval as generating a joint video–text distribution from Gaussian noise. UCoFiA [41] introduces a three-level coarse-to-fine alignment framework. DiscoVLA [34] focuses on efficient video feature extraction and proposes a parameter-efficient video-level attention module. Meanwhile, commonly used datasets (e.g., MSR-VTT [45], VATEX [39], and ActivityNet [18]) primarily describe single-subject actions and their captions, without capturing more complex scenarios such as interactions among multiple characters or variations in camera viewpoint.

Various VTR benchmarks are widely used in the field. MSR-VTT [45] provides a variety of open-domain video categories for general-purpose retrieval. VATEX [39] introduces large-scale multilingual captions (English and Chinese) to improve linguistic depth. DiDeMo [1] focuses on temporal grounding by linking descriptions to specific event boundaries within a video. However, these datasets are limited in that they consist of 10–30 second short videos aligned with one or a few literal, simple descriptive sentences, and they are largely confined to isolated, short-form visual segments. Larger datasets such as HowTo100M [23] and WebVid-2M [3] expand the scale of video–text data by leveraging instructional and web-sourced content. Yet HowTo100M is largely restricted to instructional videos, while WebVid-2M typically provides only a single descriptive caption per video despite its diverse visual domains. In contrast, our screenplay benchmark offers complex video-text data with longer durations and rich descriptions. Unlike simple captions, screenplays guide filmmaking by detailing characters' emotions, nonverbal cues, and environmental context [8]. This results in a deeper alignment between text and visual output [5, 26], bridging gaps in previous works that align visual shots with narrative structures.

2.2 Movie Scene and Narrative Understanding

Understanding movies as visual narratives attract increasing attention in computer vision. Early benchmarks explore multimodal reasoning over movie content using short clips or coarse descriptions. MovieQA [37] introduce a large-scale dataset combining clips, subtitles, scripts, and plot summaries for question answering, while MovieNet [14] expand annotation scale with labels for scenes, characters, and cinematic attributes across over a thousand films. Other datasets emphasize narrative grounding between video and text, such as MAD [35], which aligns audio descriptions with temporal segments, and Movie101 [47] which studies story centric narration and grounding. However, these datasets rely primarily on post-production text such as captions or audio descriptions, which capture only part of the structured information present in screenplays.

More recent benchmarks target long-form video reasoning. InfiniBench [2] evaluates reasoning over videos with tasks involving temporal and causal understanding, while MovieRecapsQA [33] introduces open-ended question answering

using recap videos and summaries. MovieBench [42] further provides hierarchical annotations for long-form movie generation across scenes and shots. Despite these adavances, existing benchmarks focus on narrative comprehensive or generation rather than explicitly aligning screeplay structure with visual shots. In contrast, our work studies screeplay-to-shot grounding, enabling systematic evaluation of how dialogue and visual direction in scripts correspond to cinematic video segments.

3 The *From Script to Shot* Benchmark

3.1 Task Definition

A movie can be described at multiple levels of granularity. A shot is a continuous sequence of frames from a single camera, and a scene groups consecutive shots that share continuity in action, location, or time, forming the basic narrative unit. Shot boundaries are visual, such as cuts, whereas scene boundaries are semantic, driven by shifts in narrative context. A typical movie contains hundreds to thousands of shots but only tens to hundreds of scenes.

A screenplay organizes a narrative as an ordered sequence of scenes, each combining a heading, stage directions, and dialogue. The filmed video, however, does not retain explicit scene boundaries, and only shots remain directly observable. The task is therefore to assign each shot to the screenplay scene it realizes.

Let $\mathcal{S} = \{s_1, \ldots, s_M\}$ denote the scenes in a screenplay and $\mathcal{V} = \{v_1, \ldots, v_N\}$ the shots in a video. Each shot in $\mathcal{V}$ is represented by keyframe images and optional subtitle text, while each scene in $\mathcal{S}$ is an ordered sequence of tagged text elements. The assignment $\phi : \{1, \ldots, N\} \to \{1, \ldots, M\}$ maps every shot to a screenplay scene. In the idealized case, the assigned scene indices are non-decreasing across the shot sequence, i.e., $\phi(i) \leq \phi(i')$ for all $i < i'$, and every scene is realized by at least one shot. In practice, editorial changes such as scene reordering, deletion, or addition during post-production can violate these assumptions, making the alignment strictly harder than a clean sequential mapping.

This task poses challenges beyond standard video-text retrieval and temporal grounding. Video-text retrieval matches a short clip to a single caption in a one-to-one setting, while temporal grounding localizes a moment within a video given a single query. Screenplay-to-shot alignment instead requires a sequential mapping of hundreds to thousands of shots across an entire film. Furthermore, scene text is not a homogeneous visual caption but a mix of stage directions and dialogue, and scenes that lack dialogue cannot be matched through subtitle overlap alone.

3.2 Data Sources and Preprocessing

We build our benchmark on top of MovieNet [14], which provides screenplay-video pairs for a large collection of movies but does not include scene-to-shot alignment annotations. Screenplays and their corresponding films often diverge

due to adaptation or editorial restructuring, so we validate each pair by comparing dialogue lines against subtitle transcripts and retain only those with largely consistent content and ordering. From the validated pool, we select 50 movies spanning nine decades and 16 genres, sampling to approximate the year and genre distribution of the full MovieNet collection.

Screenplays in MovieNet are provided as raw text without scene-level segmentation, and we preprocess each selected movie as follows. We apply a movie screenplay parser [4] to detect scene headings and split each screenplay into scene-level units. The parser classifies each text element into one of several categories, of which we retain four that are relevant to alignment. Each scene s_j is represented as an ordered sequence of tagged elements:

$$s_j = \{(t_k, x_k)\}_{k=1}^{L_j}, \quad t_k \in \{\texttt{H}, \texttt{N}, \texttt{C}, \texttt{D}\} \tag{1}$$

where x_k is the text content, L_j is the number of elements in scene j, and the tags denote scene heading (`H`), stage direction (`N`), character name (`C`), and dialogue (`D`).

On the video side, MovieNet provides pre-segmented shots with frame-level indices and three keyframe images per shot, with subtitle tracks available separately as timecode-based timestamps. Because MovieNet does not include frame rate metadata, direct temporal alignment between shots and subtitles is not straightforward. We leverage a subset of shot-level subtitle annotations provided by MovieNet to back-estimate the subtitle offset and frame rate for each movie, then project all subtitle entries onto the shot timeline. As a result, each shot is represented as a pair:

$$v_i = (f_i, u_i), \quad u_i = \begin{cases} u & \text{if any subtitle overlaps with } v_i \\ \emptyset & \text{otherwise} \end{cases} \tag{2}$$

where f_i denotes the keyframe images, u the subtitle string aggregated from all overlapping lines, and $u_i \in \{u, \emptyset\}$ its assignment for shot v_i.

3.3 Annotation Pipeline

Annotating scene-to-shot alignment from scratch over an entire film is prohibitively difficult. To reduce annotator burden, we generate an initial alignment by matching dialogue lines in each screenplay scene against subtitle strings assigned to shots. This dialogue-based pre-matching provides a reasonable starting point, particularly for dialogue-heavy scenes, while leaving non-dialogue scenes and ambiguous regions for human judgment.

Annotators work sequentially through the screenplay, assigning a contiguous block of shots to each scene. When a screenplay scene has no corresponding shots due to editorial deletion, the scene is marked as unmatched. Conversely, shots depicting content not present in the screenplay are assigned to the nearest contextually appropriate scene.

Each movie is annotated by a primary annotator and then independently reviewed by a validation annotator. The reviewer examines every scene-shot boundary, flagging assignments that appear inconsistent with the screenplay

Table 1: Dataset statistics (50 films). Subtitle coverage indicates the fraction of shots with ≥1 associated subtitle line. Percentage rows report per-movie statistics.

	Min	Max	Mean	Median	Total
Shots / film	363	2,791	1,069	974	53,448
Scenes / film	32	453	154	148	7,713
Shots / scene	0	366	7.4	2	–
Subtitle coverage (%)	61.5	100.0	94.6	100.0	–
Matched scenes (%)	44.4	100.0	80.7	82.7	–
Dialogue scenes (%)	35.6	90.6	60.0	60.6	–

text or visual content. Disagreements are resolved through discussion until consensus is reached. This annotation and review process is applied to all 50 movies, yielding the final set of verified alignments.

3.4 Dataset Statistics

The 50 movies contain over 55K shots in total, of which 53,448 are assigned to screenplay scenes. The rest corresponds to studio logos, title cards, and end credits with no screenplay counterpart. The screenplays comprise 7,713 scenes, of which 79.3% are matched to at least one shot, with the remainder having no corresponding shots due to editorial deletion. Each matched scene maps to 7.4 shots on average, though the distribution is heavily skewed with a median of 2 and a maximum of 366.

Of all scenes, 56.6% contain dialogue and 43.4% do not, providing near-balanced coverage of both channels for diagnostic evaluation. Tab. 1 summarizes the full distributional statistics, with detailed breakdowns by decade, genre, and scene length provided in the supplementary material.

3.5 Evaluation Metrics

We evaluate alignment quality through Scene Intersection over Union (SIoU). For each scene s_j with ground-truth shot set G_j and predicted shot set P_j, the scene-level IoU is

$$\mathrm{SIoU}_j = \frac{|G_j \cap P_j|}{|G_j \cup P_j|}. \tag{3}$$

The primary metric, mean SIoU (mSIoU), averages over all matched scenes in a movie and then over all movies:

$$\mathrm{mSIoU} = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \frac{1}{|\mathcal{M}_f|} \sum_{j \in \mathcal{M}_f} \mathrm{SIoU}_j \tag{4}$$

where $\mathcal{F}$ is the set of movies and $\mathcal{M}_f$ is the set of matched scenes in movie f. SIoU captures both the precision and recall of shot assignment for each scene, making it robust to variations in scene length.

To isolate the contributions of dialogue and visual information, we partition the matched scenes into two subsets based on the presence of dialogue elements in the screenplay. Let $\mathcal{D}_f$ and $\mathcal{N}_f$ denote the sets of dialogue and non-dialogue scenes in movie f, respectively. We report

$$\text{mSIoU}_d = \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}}\frac{1}{|\mathcal{D}_f|}\sum_{j\in\mathcal{D}_f}\text{SIoU}_j, \quad \text{mSIoU}_{nd} = \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}}\frac{1}{|\mathcal{N}_f|}\sum_{j\in\mathcal{N}_f}\text{SIoU}_j. \tag{5}$$

The gap between mSIoU_d and mSIoU_{nd} directly quantifies how much a method relies on dialogue matching versus visual grounding.

A screenplay scene may have no realized shot in the final cut (e.g., scenes dropped in editing). Such empty-ground-truth scenes (about 20% of all scenes) are excluded from the scene-IoU metrics (mSIoU, IoU_d, IoU_{nd}), which are averaged only over scenes with at least one ground-truth shot. Reported scores therefore characterize matched scenes and may modestly overstate performance on the full alignment problem, where unmatched scenes must also be detected.

As a complementary measure, we evaluate the accuracy of predicted scene boundaries using boundary F1 (BndF1). A predicted boundary is counted as a true positive if it falls within a tolerance window of δ shots from a ground-truth boundary. Precision and recall are computed over all boundary predictions, and BndF1 is their harmonic mean:

$$\text{BndF1} = \frac{2\cdot P_\delta \cdot R_\delta}{P_\delta + R_\delta}. \tag{6}$$

While mSIoU measures the overall quality of shot-to-scene assignments, BndF1 focuses specifically on boundary localization.

We also report standard retrieval metrics to assess scene-level ranking quality. For each shot v_i with ground-truth scene $\phi(i)$, we rank all scenes by the method's similarity score and record the rank r_i of the correct scene. Recall at rank k and median rank are defined as

$$\text{R@}k = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[r_i \le k], \quad \text{MedR} = \text{median}(\{r_i\}_{i=1}^{N}). \tag{7}$$

We report R@1, R@5, and MedR. These metrics complement mSIoU by measuring per-shot discrimination: a method may produce reasonable scene boundaries yet fail to rank the correct scene highly for individual shots, or vice versa.

4 Experiments

4.1 Experimental Setup

All methods in our benchmark follow a shared two-stage pipeline. In the first stage, each method computes a similarity matrix $\mathbf{S} \in \mathbb{R}^{N\times M}$ between shots and scenes using its own modality-specific scoring function. In the second stage, a common assignment algorithm converts this matrix into a monotone mapping ϕ that respects temporal ordering. We use Drop-DTW [9], a variant of

dynamic time warping that permits unmatched elements, combined with a Gaussian length prior [10, 21] that biases each shot toward a positional window proportional to the corresponding scene's text length. This design ensures that performance differences across methods reflect the quality of the similarity signal rather than the assignment strategy.

Dialogue methods compute similarity exclusively from text, matching scene dialogue against shot-level subtitle strings. BM25 [13] and TF-IDF [28] treat each scene's dialogue as a document and each shot's subtitle as a query, producing sparse lexical similarity scores via BM25 and TF-IDF cosine similarity, respectively. M/S [8] instead applies word-level DTW to align screenplay dialogue tokens to subtitle tokens, deriving boundary constraints that guide a constrained dynamic program using Jaccard word overlap as the scoring function. All three methods are entirely text-based and use no visual information. Further details are provided in the supplementary material.

Visual methods replace text matching with contrastive image-text encoders. CLIP [27] and SigLIP [48] encode the first keyframe of each shot and compute cosine similarity against sentence-level text chunks from the screenplay, max-pooled across sentences within each scene. InternVideo2 [40] operates at the video level, taking a pseudo-video of four repeated keyframes and comparing it against the concatenated scene text. All visual methods use no subtitle information.

Temporal grounding methods were originally designed to localize a single text query within a short video clip. We adapt three recent models to assess their transferability to our long-form setting. TAN [12] extracts S3D-G [44] features from a 16-frame pseudo-clip constructed by repeating each shot's keyframe, and aligns them with Word2Vec [24]-based scene text features through a multimodal Transformer. MATR [19] and T-MASS [38] both build on CLIP ViT-B/32, with MATR applying a joint Transformer encoder over concatenated video-text tokens and T-MASS using text-conditioned cross-attention to pool frame-level features into per-pair representations. All three models produce a similarity matrix that is fed into the shared assignment pipeline.

We also evaluate three video-text retrieval models as baselines: CLIP4Clip [22], DiffusionRet [16], and DiscoVLA [34]. These models are originally designed to match short video clips with descriptive captions. CLIP4Clip extends CLIP for video–text retrieval, DiffusionRet uses diffusion-based retrieval, and DiscoVLA improves video–language alignment for retrieval. To adapt them to screenplay-to-shot alignment, we treat each shot as a short video clip represented by its keyframes and encode the corresponding scene text as the query. This evaluation tests whether retrieval models trained on caption-style supervision transfer to the heterogeneous language of screenplays.

Finally, we evaluate an adaptive fusion of the dialogue and visual channels. The fused similarity for shot v_i is $(1-\alpha)\,\mathbf{S}^{\mathrm{bm25}} + \alpha\,\mathbf{S}^{\mathrm{vis}}$ when $u_i \neq \emptyset$, and $\mathbf{S}^{\mathrm{vis}}$ otherwise, with both matrices row-normalized. We sweep $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$ and report the best-performing configuration. This design tests whether the two channels provide complementary signals and whether their combination can close the gap between dialogue and non-dialogue scenes.

Table 2: Aggregate results across all movies. Best per paradigm in **bold**. Fusion weight α is selected by 5-fold movie-level cross-validation (Tab. 3), not on the test set.

Paradigm	Method	mSIoU	BndF1	IoU_d	IoU_{nd}	**R**@1	**R**@5	MedR
Dialogue	BM25 [13]	**0.399**	0.543	0.505	**0.190**	0.389	0.573	5.82
Dialogue	TF-IDF [28]	0.396	**0.546**	**0.507**	0.175	0.346	**0.576**	**5.65**
Dialogue	M/S [8]	0.324	0.480	0.412	0.156	**0.446**	0.545	6.10
Visual	Qwen3-VL-Emb. [20]	**0.500**	**0.629**	**0.521**	**0.456**	**0.257**	**0.536**	**6.40**
Visual	FG-CLIP 2 [43]	0.317	0.497	0.329	0.289	0.174	0.417	10.70
Visual	CLIP [27]	0.252	0.434	0.253	0.252	0.144	0.375	11.88
Visual	InternVideo2 [40]	0.222	0.429	0.227	0.205	0.128	0.306	19.52
Visual	SigLIP [48]	0.023	0.189	0.020	0.035	0.048	0.167	30.58
Temporal	T-MASS [38]	**0.130**	**0.300**	**0.123**	**0.141**	**0.088**	**0.261**	**21.02**
Temporal	MATR [19]	0.016	0.205	0.014	0.028	0.023	0.078	66.49
Temporal	TAN [12]	0.009	0.181	0.006	0.015	0.003	0.026	67.32
VTR	CLIP4Clip [22]	0.012	0.213	0.010	0.017	0.006	0.043	**58.57**
VTR	DiffusionRet [16]	0.011	0.217	0.008	0.016	**0.014**	0.059	67.30
VTR	DiscoVLA [34]	**0.013**	**0.221**	**0.011**	**0.026**	0.013	**0.059**	69.80
Fusion(α=0.7)	Qwen3-VL-Emb.&BM25	**0.599**	**0.695**	**0.654**	**0.496**	**0.448**	**0.686**	**2.70**
Fusion(α=0.5)	CLIP&BM25	0.531	0.633	0.597	0.403	0.416	0.594	4.50
Fusion(α=0.5)	FG-CLIP 2&BM25	0.526	0.638	0.598	0.378	0.426	0.606	4.30

4.2 Benchmark Results

Tab. 2 summarizes the results across all paradigms. Among dialogue methods, BM25 [13] and TF-IDF [28] perform comparably (0.399 and 0.396 mSIoU), both outperforming M/S [8] (0.324). All three exhibit the same structural pattern: strong alignment on dialogue scenes ($\text{IoU}_d \approx 0.50$) but sharp degradation on non-dialogue scenes ($\text{IoU}_{nd} \leq 0.19$), confirming that subtitle matching alone cannot ground scenes that lack spoken content.

Visual encoders score lower than dialogue methods overall but behave differently across scene types. Qwen3-VL-Embedding [20] is the strongest visual encoder (0.500 mSIoU), followed by FG-CLIP 2 [43] (0.317) and CLIP (0.252), while SigLIP [48] largely fails (0.023). Notably, CLIP achieves nearly identical performance on dialogue and non-dialogue scenes (0.253 vs. 0.252), indicating that visual similarity is not dependent on the presence of dialogue. Temporal grounding and VTR models perform worst: T-MASS [38] retains 0.130 mSIoU, whereas MATR [19], TAN [12], and the VTR baselines fall below 0.02, suggesting that models trained to localize a single query within a short clip do not transfer to full-length film alignment without architectural adaptation.

Adaptive fusion is the strongest paradigm overall. Qwen3-VL-Embedding [20] fused with BM25 leads on every metric (0.599 mSIoU, 0.695 BndF1), a relative gain of 50% over BM25 alone, and even the widely used CLIP, when fused with BM25, reaches 0.531 mSIoU. Taking CLIP and BM25 as a controlled reference, the fusion improves both channels: IoU_d rises from 0.505 (BM25 alone) to 0.597,

Table 3: 5-fold cross-validation for α selection (mSIoU). n/5: number of folds in which this α was selected by validation.

Encoder (α, n/5)	f-1	f-2	f-3	f-4	f-5	total
CLIP (0.5, 5/5)	0.469	0.615	0.534	0.452	0.586	0.531
FG-CLIP 2 (0.5, 5/5)	0.472	0.630	0.517	0.467	0.548	0.527
Qwen3-VL-Emb. (0.5, 1/5)	0.541	**0.652**	0.611	0.540	0.612	0.591
Qwen3-VL-Emb. (0.7, 4/5)	**0.547**	0.636	**0.632**	**0.567**	**0.614**	**0.599**

and IoU_{nd} rises from 0.252 (CLIP alone) to 0.403. The gap between dialogue and non-dialogue scenes narrows from 0.315 to 0.194, yet remains substantial, pointing to a persistent visual encoding bottleneck that even the strongest encoder does not resolve.

4.3 Analysis of Alignment Results

We select the fusion weight α by 5-fold, movie-level cross-validation, never on the test set. For CLIP + BM25 the per-fold optimum is unanimous (α=0.5 in all five folds) and the cross-validated test mSIoU equals the full-sweep value (0.531); for Qwen3-VL-Embedding + BM25, α=0.7 is selected in four of five folds and the cross-validated test mSIoU is 0.599, confirming that the gain is not an artefact of test-set tuning. The optimal α rises with encoder strength (0.5 for CLIP and FG-CLIP 2 [43], 0.7 for Qwen3-VL-Embedding), consistent with stronger visual encoders warranting a larger visual weight in fusion.

The fusion gain on non-dialogue scenes is the central finding. CLIP [27] alone achieves only 0.252 IoU_{nd}, yet fusing it with BM25 [13] raises this to 0.403, far beyond what either channel reaches independently, even though the same CLIP features are used in both settings. The improvement does not come from better visual encoding. Rather, BM25 accurately anchors dialogue scenes to their correct positions, and the temporal ordering constraint in Drop-DTW propagates these anchors to neighboring non-dialogue scenes, narrowing the search space for visual matching. We refer to this as the *dialogue anchoring effect*: reliable dialogue alignment indirectly constrains the placement of non-dialogue scenes that would otherwise be ambiguous.

Two structural limits of current encoders follow. First, CLIP's near-identical dialogue/non-dialogue scores (0.253 vs. 0.252), despite dialogue scenes containing substantially more text, show that contrastive encoders are blind to the distinction between dialogue and stage directions: including or excluding dialogue does not change the visual matching signal. The bottleneck therefore lies not in the amount of text available but in the encoder's inability to interpret screenplay-specific language such as stage directions and scene headings. Second, the spread among temporal grounding models (0.130 for T-MASS vs. 0.009 for TAN) reflects a scale mismatch beyond domain shift: all three were designed to localize a single query within a clip of a few minutes, whereas our task requires simultaneous alignment of hundreds of scenes across a two-hour film. This identifies input-length scalability as a concrete design requirement for future models.

Table 4: Application to movie scene segmentation. (a) Performance on MovieNet-SSeg without any task-specific training. (b) Boundary agreement between screenplay-shot alignment and scene segmentation annotations across 44 overlapping movies.

Method	Type	F1	mIoU
GraphCut [30]	Unsup.	–	20.70
DP [11]	Unsup.	–	32.00
Grouping [32]	Unsup.	–	37.20
Ours (alignment)	Zero-shot	25.61	32.24
LGSS [29]	Sup.	–	48.80
BaSSL [25]	Sup.	47.02	50.69
TranS4mer [15]	Sup.	48.36	51.91
CAT [46]	Sup.	51.93	53.67

(a) MovieNet-SSeg performance.

Metric	Value
Exact-match F1	0.477
Tolerance ±3 F1	0.752
Tolerance ±5 F1	0.813
Cohen's κ	0.425
Alignment / Seg. boundaries	1.37×
Mean distance (Align→Seg)	4.6 shots
Mean distance (Seg→Align)	2.9 shots
Ambiguous near Al (± 2)	77.5%

(b) Boundary agreement (44 movies).

Together, these findings point to two actionable directions. First, visual encoders need to move beyond object-centric caption matching toward understanding screenplay language, where stage directions describe actions, camera movements, and spatial relationships rather than object categories. Second, the dialogue anchoring effect suggests that future architectures could explicitly model the interplay between dialogue-grounded and visually-grounded scenes, rather than treating each scene independently.

5 Downstream Applications

5.1 Zero-Shot Movie Scene Segmentation

Scene segmentation is a core task in movie understanding that partitions a film into semantically coherent scenes. Existing methods typically train boundary detectors on visual or audiovisual features without access to screenplay structure. Our alignment provides a complementary signal: if a model can ground screenplay scenes in video, the resulting assignment ϕ directly induces scene boundaries at shots where $\phi(i) \neq \phi(i{+}1)$.

Without any task-specific training, this zero-shot baseline achieves 32.24 mIoU on MovieNet-SSeg, as shown in Tab. 4a, matching the classical unsupervised method DP (32.00 mIoU). The gap to supervised methods (50–54 mIoU) is expected, as our alignment optimizes scene identity rather than boundary detection.

To understand this gap, we compare alignment-derived boundaries against segmentation annotations on the 44 movies shared between our benchmark and MovieNet [14]. The results are summarized in Tab. 4b. Exact-match F1 is 0.477, but allowing a ±3-shot tolerance raises it to 0.752, indicating that the two tasks identify similar structural boundaries with a small positional offset. Alignment produces 1.37× more boundaries on average, because the screenplay treats short

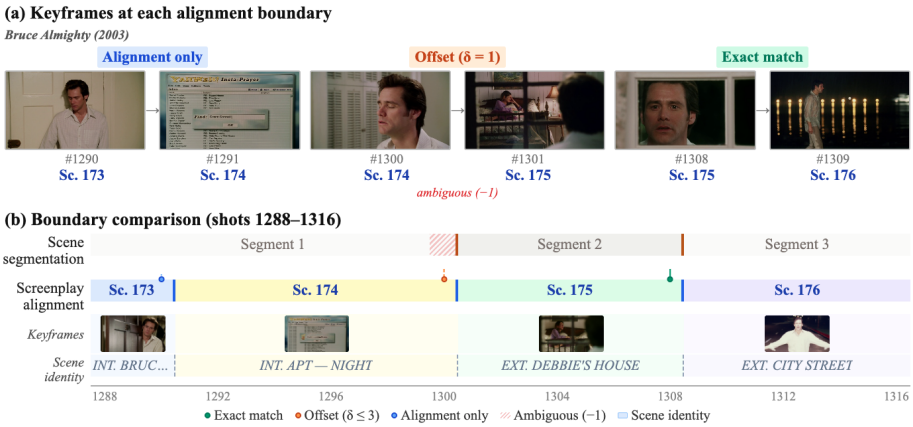

Fig. 2: Alignment vs. segmentation boundaries on *Bruce Almighty* (shots 1288–1316). (a) Keyframes at each alignment boundary: • alignment-only (same apartment, different screenplay scenes), • 1-shot offset (annotator marked as ambiguous), • exact match. (b) Timeline: segmentation yields 3 segments; alignment yields 4 scenes with heading-level identity (bottom row), semantic information absent from boundary labels alone.

transitions such as cutaways and inserts as independent scenes, whereas segmentation annotators tend to absorb them into neighboring segments. Notably, 77.5% of shots marked as ambiguous (−1) by segmentation annotators fall within ±2 shots of an alignment boundary, suggesting that boundary ambiguity in visual annotation coincides with screenplay-level scene transitions.

Fig. 2 illustrates this on a representative excerpt from *Bruce Almighty*. Segmentation identifies three segments, while alignment identifies four screenplay scenes with three boundaries. At shot 1308, the two methods agree exactly (•); at shot 1300 both detect a transition but alignment places it one shot earlier (•, δ=1), and the annotator also flagged this shot as ambiguous (−1). At shot 1290, alignment marks a scene change between two scenes set in the same apartment (•), which segmentation merges. Crucially, alignment also provides *scene identity*: each segment is linked to a screenplay scene with heading and dialogue content, offering semantic grounding absent from boundary labels alone.

These findings confirm that screenplay-shot alignment captures scene structure that overlaps with, but is finer than, segmentation. This makes alignment a practical source of zero-shot pseudo-labels for training segmentation models on unannotated films, and provides richer semantic grounding than boundary labels alone.

5.2 Screenplay-Guided Video Generation

Screenplays encode narrative intent in a format designed for film professionals, such as slug lines, action descriptions, and character cues, rather than the explicit visual language that text-to-video models expect. This format gap means that

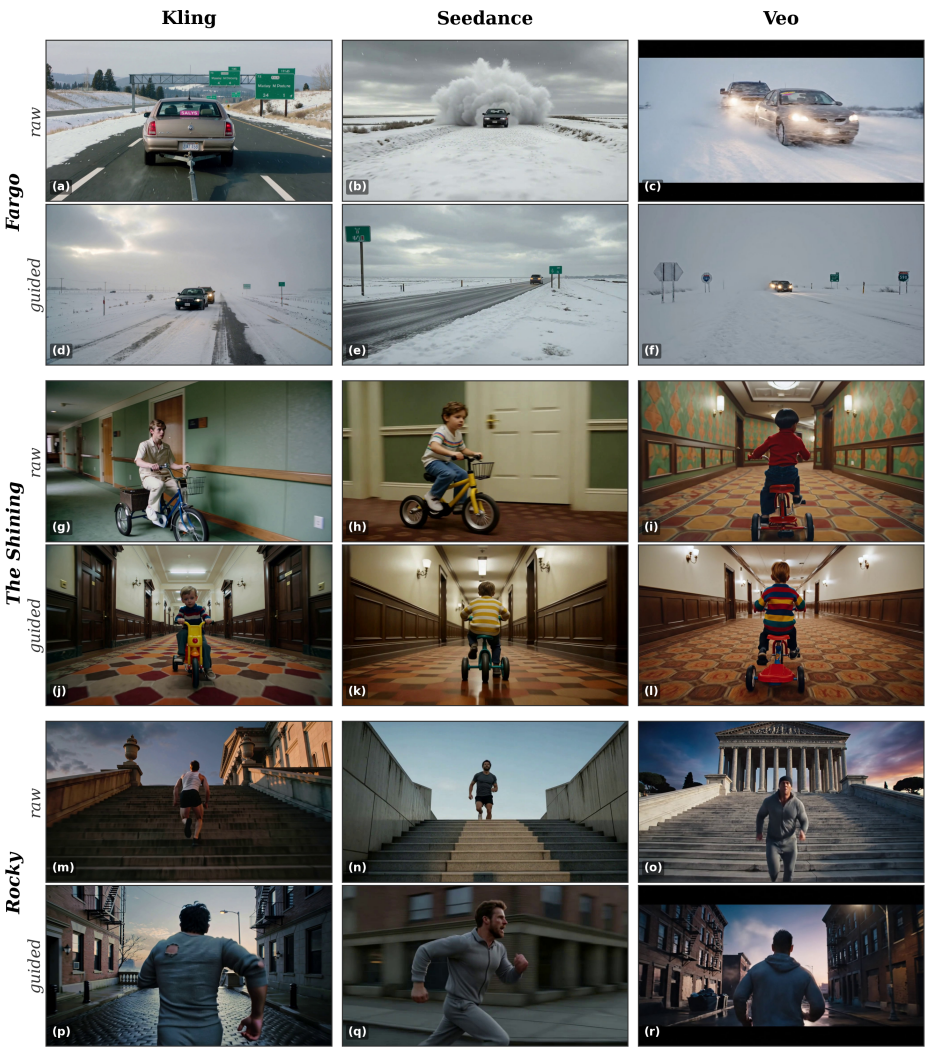

Fig. 3: Text-to-video generation from raw screenplay text vs. alignment-guided prompts across three services (Kling, Seedance, Veo) and three scenes. Raw text retains screenplay conventions (slug lines, camera shorthand, character names) that lack explicit visual detail; guided prompts, informed by our screenplay–shot alignment, inject concrete cinematographic cues such as lighting, set design, color palette, and spatial composition. The guided condition consistently produces frames closer to each film's visual identity.

feeding raw screenplay text into services such as Kling, Seedance, or Veo yields frames that lack the cinematic specificity of the source film.

Our screenplay–shot alignment directly addresses this gap. Each aligned pair associates a screenplay scene with visual descriptions derived from its corre-

sponding keyframes, providing concrete examples of how textual narrative was actually realised on screen. We use these pairs as few-shot demonstrations for an LLM, which rewrites a target screenplay scene into a visually grounded prompt enriched with lighting, set design, color palette, and spatial composition, details that the raw text never makes explicit.

We evaluate on three scenes spanning distinct visual registers, Fargo (opening drive), The Shining (Danny's corridor ride), and Rocky (training staircase), generating videos under both raw and guided conditions across three commercial services (18 videos total). Fig. 3 shows representative frames.

Raw inputs produce systematic errors: models misinterpret screenplay shorthand, rendering Danny as an adult (g), literalizing metaphorical language as a snow explosion (b), or hallucinating a Greco-Roman temple for "art museum stairs" (o). Guided prompts consistently recover film-specific visual identity across all three services, including the Overlook's hexagonal carpet and child-scale tricycle (j–l), Fargo's bleak Midwestern flatness (d–f), and Rocky's neoclassical museum steps at dawn (p–r). The contrast demonstrates that our alignment benchmark provides the paired training signal needed to automate the translation from screenplay conventions to cinematic visual language.

6 Conclusion

Grounding a screenplay in its filmed realization differs fundamentally from matching captions to clips: a screenplay mixes dialogue and stage directions rather than a direct visual description, so neither text overlap nor visual similarity alone suffices, and the strongest alignment emerges only when the two channels constrain each other through temporal structure.

Evaluating fourteen methods on our benchmark exposes two bottlenecks—visual encoders trained on object-centric captions cannot interpret screenplay language, and temporal grounding models built for short clips do not transfer to full-length films—while the dialogue anchoring effect makes fusion the strongest paradigm.

Two downstream applications, zero-shot scene segmentation and screenplay-guided video generation, confirm that the alignment signal generalizes beyond the benchmark.

Limitations. Our benchmark derives from a single source (MovieNet), skewing toward English-language Western cinema, and assumes a corresponding screenplay is available—a deliberate scope choice distinct from screenplay-free video understanding. Alignment is also evaluated only on matched scenes (Sec. 3.5); jointly detecting unmatched scenes is left to future work.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (RS-2020-II201361). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (RS-2024-00338439). This work was supported by the InnoCORE program of the Ministry of Science and ICT (N10250156).

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2017)
2. Ataallah, K., Bakr, E.M., Ahmed, M., Gou, C., Pahwa, K., Ding, J., Elhoseiny, M.: Infinibench: A benchmark for large multi-modal models in long-form movies and tv shows. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2025)
3. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
4. Baruah, S., Narayanan, S.: Character coreference resolution in movie screenplays. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2023)
5. Block, B.: The Visual Story: Creating the Visual Structure of Film, TV, and Digital Media. Routledge, London and New York, 3rd edn. (2020). https://doi.org/10.4324/9781315794839
6. Carreira, J., Zisserman, A.: 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)
7. Chen, T.S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.w., Jeon, B.E., Fang, Y., Lee, H.Y., Ren, J., Yang, M.H., et al.: Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
8. Cour, T., Jordan, C., Miltsakaki, E., Taskar, B.: Movie/script: Alignment and parsing of video and text transcription. In: Proceedings of the European Conference on Computer Vision (ECCV) (2008)
9. Dvornik, M., Hadji, I., Derpanis, K.G., Garg, A., Jepson, A.: Drop-dtw: Aligning common signal between sequences while dropping outliers. Proceedings of the Neural Information Processing Systems (NeurIPS) (2021)
10. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2017)
11. Han, B., Wu, W.: Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In: 2011 IEEE International conference on multimedia and expo. pp. 1–6. IEEE (2011)
12. Han, T., Xie, W., Zisserman, A.: Temporal alignment networks for long-term video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
13. Harman, D.K.: Overview of the third text retrieval conference (TREC-3). DIANE Publishing (1995)
14. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
15. Islam, M.M., Hasan, M., Athrey, K.S., Braskich, T., Bertasius, G.: Efficient movie scene detection using state-space transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
16. Jin, P., Li, H., Cheng, Z., Li, K., Ji, X., Liu, C., Yuan, L., Chen, J.: Diffusionret: Generative text-video retrieval with diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)

17. Katz, E.: Ephraim katz's the film encyclopedia (1979)
18. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2017)
19. Kumar, Y., Agarwal, U., Gupta, M., Mishra, A.: Aligning moments in time using video queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
20. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)
21. Li, Z., Chen, Q., Han, T., Zhang, Y., Wang, Y., Xie, W.: Multi-sentence grounding for long-term instructional video. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
22. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
23. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
24. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the Neural Information Processing Systems (NeurIPS) (2013)
25. Mun, J., Shin, M., Han, G., Lee, S., Ha, S., Lee, J., Kim, E.S.: Bassl: Boundary-aware self-supervised learning for video scene segmentation. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2022)
26. Papalampidi, P., Keller, F., Frermann, L., Lapata, M.: Screenplay summarization using latent narrative structure. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2020)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML) (2021)
28. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning (2003)
29. Rao, A., Xu, L., Xiong, Y., Xu, G., Huang, Q., Zhou, B., Lin, D.: A local-to-global approach to multi-modal movie scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
30. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. IEEE transactions on Multimedia (2005)
31. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
32. Rotman, D., Porat, D., Ashour, G.: Robust and efficient video scene detection using optimal sequential grouping. In: 2016 IEEE international symposium on multimedia (ISM) (2016)
33. Shaar, S., Thymes, B., Chaixanien, S., Cardie, C., Hariharan, B.: Movierecapsqa: A multimodal open-ended video question-answering benchmark. arXiv preprint arXiv:2601.02536 (2026)

34. Shen, L., Gong, G., Hao, T., He, T., Zhang, Y., Liu, P., Zhao, S., Han, J., Ding, G.: Discovla: Discrepancy reduction in vision, language, and alignment for parameter-efficient video-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
35. Soldan, M., Pardo, A., Alcázar, J.L., Caba, F., Zhao, C., Giancola, S., Ghanem, B.: Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
36. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
37. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
38. Wang, J., Sun, G., Wang, P., Liu, D., Dianat, S., Rabbani, M., Rao, R., Tao, Z.: Text is mass: Modeling as stochastic embedding for text-video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
39. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
40. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
41. Wang, Z., Sung, Y.L., Cheng, F., Bertasius, G., Bansal, M.: Unified coarse-to-fine alignment for video-text retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
42. Wu, W., Liu, M., Zhu, Z., Xia, X., Feng, H., Wang, W., Lin, K.Q., Shen, C., Shou, M.Z.: Moviebench: A hierarchical movie level dataset for long video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
43. Xie, C., Wang, B., Kong, F., Li, J., Liang, D., Ao, J., Leng, D., Yin, Y.: Fg-clip 2: A bilingual fine-grained vision-language alignment model. arXiv preprint arXiv:2510.10921 (2025)
44. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
45. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
46. Yang, Y., Huang, Y., Guo, W., Xu, B., Xia, D.: Towards global video scene segmentation with context-aware transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2023)
47. Yue, Z., Zhang, Q., Hu, A., Zhang, L., Wang, Z., Jin, Q.: Movie101: A new movie understanding benchmark. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2023)
48. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)