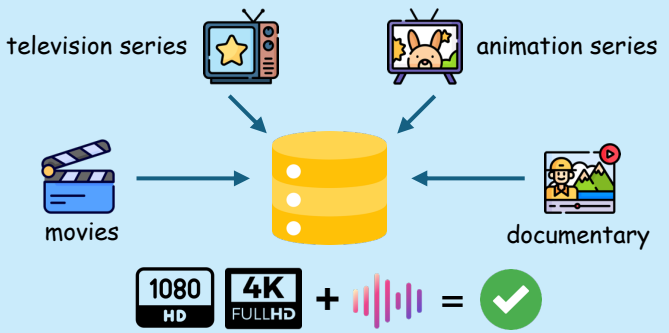
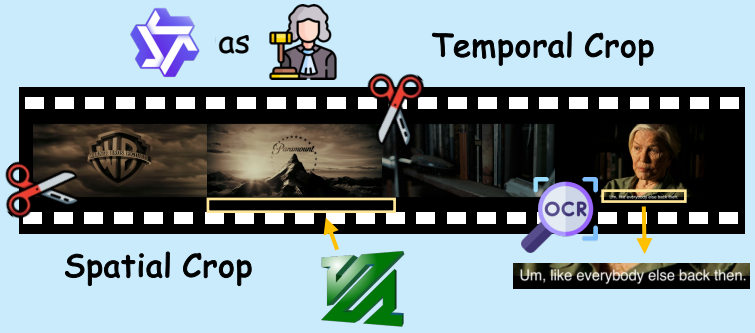
Sec 3.1 Data Curation & Quality Control

Data Collecting



Spatiotemporal Crop



Quality Assessment

Visual Quality: (1) Aesthetic Score. (2) Technical Score. (3) Motion Smoothness.

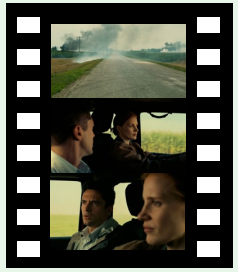
Audio Quality: (1) DNSMOS for low-level fidelity. (2) Temporal variance of CLAP embeddings for semantic diversity

Audio-Visual Alignment: (1) ImageBind cosine similarity. (2) SyncNet for lip-sync precision

Sec 3.2 Narrative Sequence Parsing

TransNetV2

#shot no	start_t	end_t	start_f	end_f
0	0.00	3.00	0	72
1	3.00	8.00	73	192
2	8.00	11.00	193	264
...	...	...	...	...



Parsing Engine (Qwen3.5-27B)

context-aware sliding window

Parse when the state breaks with unexplainable reasons. ex: background and characters all change suddenly

Merge under the following four conditions (synthesized from Christian Metz's eight Syntagmas)

① multi-angle & spatial coherence: same scene, varying viewpoints

② cross-cutting & narrative insertions: alternating spatial or temporal states, shared narrative tension

③ causal action & ellipsis: spatiotemporal leaps due to explainable cause

④ Montage: spatiotemporal compression, thematic juxtaposition

Anti-Over-Parsing: Apply ≧ 20 pacing prior to prevent narrative fragmentation.

Sec 3.3 Structured Annotation

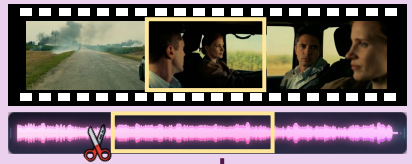
Visual Captioning (Qwen3.5-35B-A3B)

Global:
<char1>: A woman with fair skin and red hair ... (24 words)
<char2>: A man with short dark hair, wearing ... (20 words)
<scene1>: A cracked, empty asphalt road which ... (36 words)
<scene2>: A dark, compact, utilitarian car ... (27 words)

Shot-Level:
shot_type: OTS | RVS | Static | EL | Short
shot_desc: An OTS shot from behind a male ... (23 words)
transition_type: cut
transition_desc: A straight cut flips to a tighter... (22 words)
brief prompt: <char1> drives the car, looking ... (11 words)
detailed prompt: A hard cut brings the scene from the smoky exterior road into the dark interior of ... (132 words)

Audio-Visual binding (Qwen3-Omni-30B-A3B)

Leverage multimodal information



no asr, cut it!
*<think> **2.71s-4.61s** someone says <S> **His solution was correct.** .....<E>, from **audio** it seems like a woman. From **video** I see the woman's lip move. With **visual and audio description**, it must be <char1> which is speaking here! <think>*

subtask1: sentence-level ASR
[1.08~2.41] You're absolutely positive?
[2.71~4.61] His solution was correct. He'd had it for years.
.....

subtask2: shot-level audio prompt
audio_prompt: Muted car interior ambience with a steady engine hum, soft road rumble, faint cabin vibration ... (30 words)

subtask3: subjects voice description
<char1>: Her voice is calm, steady, and has a medium... (21 words)
<char2>: His voice is of a medium pitch, and he ... (25 words)

Audio Captioning (Qwen3-Omni-30B-A3B)