ONE VIDEO, MANY VIEWERS: A BENCHMARK FOR PERSONA-CONDITIONED VIDEO HIGHLIGHTS AND CHAPTER TITLES

Anonymous authors
Paper under double-blind review

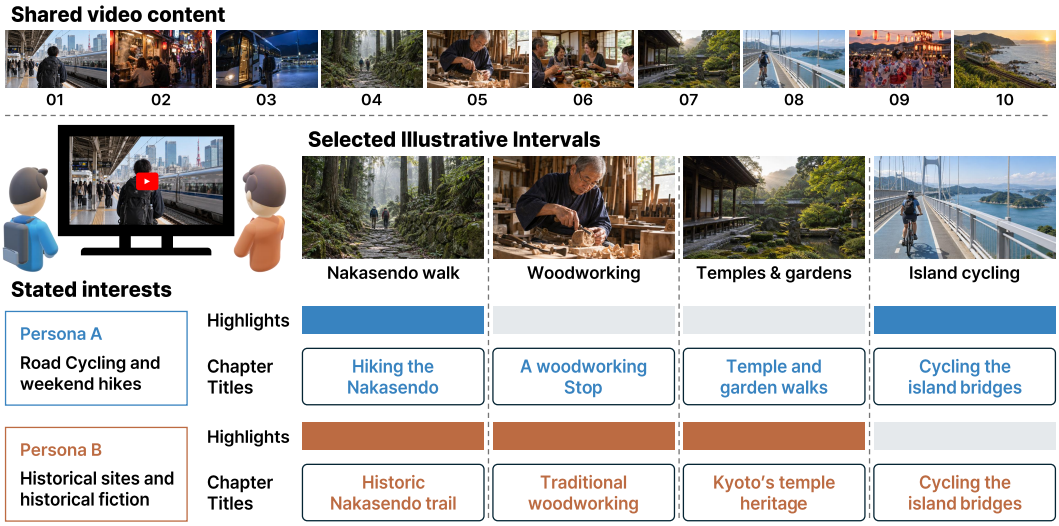


Figure 1: One video, many viewers. An illustrative comparison of highlights and chapter titles for two personas on the same video, with supplied intervals held fixed.

ABSTRACT

Highlights and chapters help viewers navigate long videos, yet they are usually produced once and shared by all viewers. Which moments matter and how each part should be described can differ between viewers, but existing evaluations rarely test whether these outputs fit a particular viewer. We study highlight selection and chapter titling conditioned on a viewer persona, a short description of a viewer's background and interests. Evaluating this setting requires annotations for many personas on the same video. Large language models make such annotations easy to produce, but differences between annotations may reflect run-to-run variation rather than the persona. We introduce *One Video, Many Viewers*, a benchmark with synthetic annotations for 259 source videos across highlight selection and chapter titling, built with 24 census-grounded synthetic personas. Highlight annotations are guided by a video-specific perspective derived from the persona, and chapter titles are generated for fixed intervals. Repeated-generation diagnostics show that persona-associated differences recur within an annotation model, whereas persona-specific title contrasts are much smaller across models than within each model. Evaluating existing models without benchmark-specific training, we find small but consistent gains over another persona in base AP for highlight models that score every unit. In contrast, we find no consistent gains for models that select units or for chapter titling. These results show that agreement with persona-conditioned references should be read alongside comparisons with other-persona and persona-free conditions. We release the annotations, personas, and perspectives for viewer-specific evaluation.

1 INTRODUCTION

Highlight detection and video chaptering help viewers find relevant content without watching an entire video. Highlight detection (Song et al., 2015; Sun et al., 2014; Sul et al., 2023) identifies

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.

salient moments, while chaptering (Yang et al., 2023a; Ventura et al., 2025) organizes the video into titled segments that support navigation. Although viewers may differ in what they find relevant, these studies evaluate highlights and chapters without explicitly assessing whether these outputs align with individual viewers' interests.

Prior work incorporates viewer preferences into highlight detection through two main approaches. Query-based highlighting finds moments relevant to a natural-language query that specifies what the viewer wants to find (Lei et al., 2021), and personalized highlight detection infers preferences from previously selected highlights or viewing histories (García del Molino & Gygli, 2018; Chen et al., 2021; Lee et al., 2025).

However, a query must be written for every video, and viewing histories offer little guidance when a video covers topics absent from a viewer's past interactions. We thus study highlight selection and chapter titling conditioned on a viewer persona, a short textual description of the viewer's background and interests that can be reused across videos. Unlike a query, a persona does not specify what to look for in a particular video, so a model must connect the stated interests to the video content to select highlights and to title supplied intervals. For example, Figure 1 illustrates how viewers interested in outdoor activities or in history may prefer different highlights and chapter titles for the same video.

Evaluating this setting requires annotations for multiple personas on the same video. Large language models can construct them at scale, but how much the persona actually changes their outputs must be measured (Hu & Collier, 2024; Schäfer et al., 2025), since an annotation may appear reasonable for its persona while suiting others equally well. Because outputs also vary across repeated runs, persona effects must be separated from this variation by testing whether annotations for the same persona are more similar to one another than to those for different personas.

We introduce *One Video, Many Viewers*, a benchmark with synthetic reference annotations for 165 source videos in highlight selection and 94 in chapter titling, built with 24 synthetic personas derived from census-grounded profiles. Highlight annotation first elicits a *perspective*, a short, video-specific description of what to prioritize given the persona's interests, whereas chapter titles for supplied intervals are generated directly from the persona and the speech in each interval. Pairing each video with multiple personas allows us to compare persona conditions for the same video.

We examine whether persona-conditioned annotations remain consistent across repeated runs and across annotation models, the language models that generate the references. Within each annotation model, annotations for the same persona are more similar than those for different personas. For chapter titles, however, this same-persona advantage is much smaller when the annotations come from different models. These results suggest that persona-associated differences are reproducible within a model, but how a persona shapes chapter titles depends on the model.

Using these references, we evaluate existing models without benchmark-specific training and test whether providing the target persona improves agreement over providing a different persona or no persona. For highlight selection, models that score each video unit show small but consistent gains in average precision (AP) on the base highlight set when given the target persona rather than another persona. Models that directly output a set of selected units show no consistent gain. For chapter titling, the target persona yields no consistent gain, and its effects depend on the annotation model used to generate the references. High agreement with target-persona references does not by itself demonstrate that a model uses the persona and should be interpreted along with results from other-persona or persona-free conditions.

Our contributions are threefold.

- **Task and benchmark.** We formulate persona-conditioned highlight selection and chapter titling over supplied intervals and release *One Video, Many Viewers*, whose synthetic references support comparisons across personas on fixed video content.
- **Annotation diagnostics.** We compare same- and other-persona agreement within and across annotation models and with and without perspectives, showing that persona differences repeat within each model, whereas the two models share little persona-specific title wording.
- **Controlled evaluation.** We separate reference agreement from the benefit of the persona through other-persona and persona-free controls, finding consistent gains over another persona in base AP for highlight models that score every unit.

2 Related Work

2.1 Video Highlight Detection and Chaptering

Highlight detection ranks or selects video segments by saliency, typically learned from supervision shared across viewers, such as edited videos, video-level categories, or annotated highlights (Sun et al., 2014; Hong et al., 2020; Badamdorj et al., 2021). Query-conditioned highlight detection instead makes this selection depend on a natural-language description of the content of interest. QVHighlights (Lei et al., 2021) provides a benchmark for this setting, with annotations for query-relevant moments and saliency scores for query-relevant clips. QD-DETR (Moon et al., 2023b) strengthens the connection between a query and the selected content through query-dependent video representations. These approaches evaluate selection under an explicit content request.

Video chaptering, on the other hand, organizes a video into temporal segments and assigns a descriptive title to each (Cao et al., 2022; Pu et al., 2025). VidChapters-7M (Yang et al., 2023a) provides a large-scale benchmark built from user-provided chapter annotations, supporting chapter generation both with and without supplied temporal boundaries. Chapter-Llama (Ventura et al., 2025) uses timestamped speech transcripts and frame captions to generate chapter boundaries and titles with a large language model. Our benchmark evaluates highlight selection and chapter titling under an explicit viewer-persona condition, holding supplied intervals fixed for titling.

2.2 Conditioning Signals for Personalization

Beyond explicit content requests, personalized highlight detection infers a viewer's preferences from past behavior, such as previously selected highlights or created GIFs (García del Molino & Gygli, 2018; Rochan et al., 2020; Chen et al., 2021). More recently, HIPPO-Video (Lee et al., 2025) uses an LLM-based user simulator to construct watch histories and evaluates history-conditioned segment saliency prediction. Whereas queries specify requested content and histories provide examples of past behavior, we condition highlight selection on an explicit description of the viewer's background and interests.

Personalized text generation conditions outputs on information about the intended reader or the author. PENS (Ao et al., 2021) generates news headlines tailored to a reader's interests inferred from behavioral data. LaMP (Salemi et al., 2024) evaluates personalized language understanding and generation using users' historical inputs and outputs, including title generation conditioned on an author's previous article–title pairs. We instead use synthetic persona descriptions to specify the intended audience for chapter titles.

2.3 Evaluating Synthetic Annotations

Language models generate text annotations evaluated against human labels (Gilardi et al., 2023) and simulate survey responses conditioned on sociodemographic descriptions (Argyle et al., 2023). However, model opinions can remain misaligned with human groups even under explicit demographic conditioning (Santurkar et al., 2023). Conditioning on such descriptions does not by itself establish fidelity to human judgments.

In personalized video highlighting, HIPPO-Video (Lee et al., 2025) uses human evaluation to assess query plausibility, video-selection agreement, and clip–preference alignment. It also uses automated methods to distinguish simulated watch histories from real ones. However, these evaluations do not establish whether differences between preference conditions exceed variation across repeated annotation runs. Our benchmark complements them by comparing annotation agreement for the same and different personas, both within and across annotation models and with and without a generated perspective. These diagnostics assess consistency under persona conditioning and do not replace human evaluation of annotation quality.

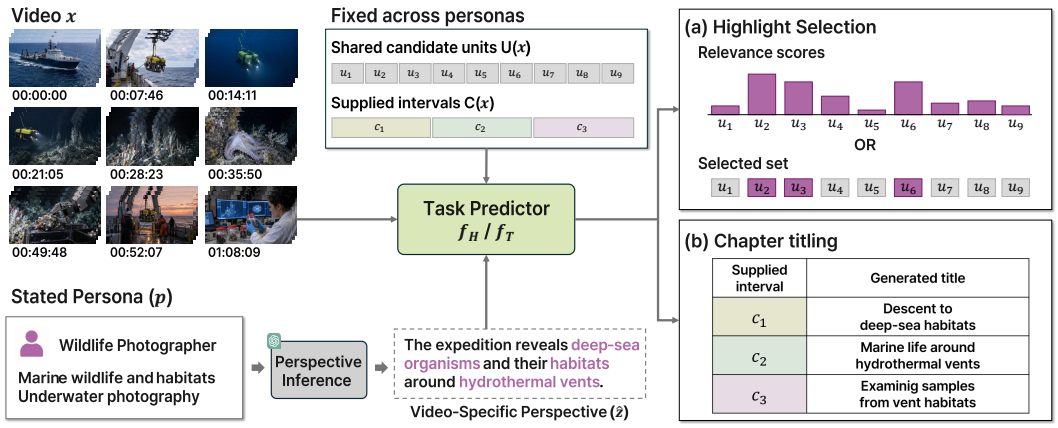


Figure 2: Task overview. A model maps a video and a stated persona to unit scores or a selected set for highlights and to titles for supplied intervals, optionally through an inferred perspective.

3 Task Formulation

3.1 Viewer Personas

We study video highlight selection and chapter titling conditioned on a stated viewer persona. Let x denote a video, whose content comprises the visual and spoken information, and let p denote a textual description of a viewer's background and interests. The objective is to select or emphasize content that is supported by x and relevant to the interests expressed in the viewer persona p. The persona specifies the intended audience and guides which content receives attention. It does not supply facts or a writing style. Pairing each video with multiple personas allows us to compare persona conditions for the same video. A persona may still admit multiple appropriate outputs, and personas with overlapping interests may share them.

We evaluate model outputs against persona-conditioned synthetic annotations, which serve as reference annotations and are described in Sections 4.2 and 4.4. Agreement with these references does not by itself establish that an output is supported by the video or relevant to the persona.

3.2 Task Definitions

We define two tasks, highlight selection and chapter titling, and evaluate them separately. In both, outputs are defined on a temporal structure that depends only on the video, so all personas are compared on the same units or intervals. Figure 2 illustrates both tasks. A model is any method that maps the task inputs to the required output, whether a single network or a multi-stage pipeline. For a fixed video x, we index outputs by the persona p, and the argument after the semicolon in $f(x, p; \cdot)$ denotes the fixed temporal structure.

For highlight selection, $\mathcal{U}(x) = (u_1, \ldots, u_M)$ denotes the candidate units of x, where M is the number of units. Each unit $u_i = [a_i, b_i)$ spans from a_i to b_i seconds, and the units are consecutive and non-overlapping. Given x and p, a model either scores every unit or selects a subset of units:

$$\widehat{\boldsymbol{s}}_p = f_H^{\text{score}}(x, p; \mathcal{U}(x)) \in \mathbb{R}^M, \tag{1}$$

$$\widehat{H}_p = f_H^{\text{set}}(x, p; \mathcal{U}(x)) \subseteq \{1, \ldots, M\}, \tag{2}$$

where $\widehat{s}_{p,i}$ is the predicted relevance of u_i to p and $\widehat{H}_p$ contains the indices of the selected units. Each model produces one of these two outputs, and models that return time spans are mapped to the units they overlap. We evaluate score-induced rankings with AP and selected sets with F1, and Section 5.1 specifies which units in the reference annotations count as relevant.

For chapter titling, x is accompanied by supplied intervals $\mathcal{C}(x) = (c_1, \ldots, c_K)$ ordered by start time, where K is the number of intervals and each $c_k = [s_k, e_k)$ spans from s_k to e_k seconds. The intervals are derived from creator-provided chapters and need not cover the entire video. Given x and p, a model generates one title per interval:

$$\widehat{T}_p = f_T(x, p; \mathcal{C}(x)) = (\widehat{t}_{p,1}, \ldots, \widehat{t}_{p,K}), \tag{3}$$

where $\widehat{t}_{p,k}$ is a natural-language title for c_k. Each title should be supported by the content of x within c_k, denoted $x|_{c_k}$, and emphasize the content most relevant to p. Holding the intervals fixed separates title generation from boundary prediction, the estimation of s_k and e_k, without assuming that different viewers would choose identical boundaries.

3.3 From Personas to Perspectives

A persona p describes a viewer's background and interests independently of any particular video, so it does not specify which content of x to emphasize. A perspective z makes this connection explicit through a short textual account of content in x that may be relevant to the persona for a given task. It represents an interpretation of the persona in context, not an observed viewer request.

A model may infer a perspective at test time and use it as an optional intermediate. Let g_H and g_T denote perspective-generation functions for highlight selection and chapter titling, respectively:

$$\widehat{z}_H = g_H(x, p), \qquad \widehat{z}_{T,k} = g_T(x|_{c_k}, p), \tag{4}$$

where $\widehat{z}_H$ is a highlight perspective for the whole video and $\widehat{z}_{T,k}$ is a title perspective for interval c_k that uses only content within that interval. As in Section 3.2, $\hat{\cdot}$ denotes model outputs. The prediction functions may then take the inferred perspective in place of, or alongside, the persona while retaining access to the video content. Models may also predict directly from x and p.

A perspective z^{ref}, where ref indicates its use in reference annotation construction, has a different role. Our highlight references are constructed through such perspectives, whereas our title references are generated directly from the persona. Supplying z^{ref} to a model would reveal information used to construct its references, so all evaluated models receive only the inputs defined in Section 3.2.

4 Benchmark Construction

4.1 Video Sources and Viewer Personas

We use QVHighlights (Lei et al., 2021) clips for highlight selection and a video subset distributed with Chapter-Llama (Ventura et al., 2025) for chapter titling. QVHighlights provides query-conditioned saliency annotations, and the Chapter-Llama subset provides automatic speech recognition (ASR) transcripts and creator chapters, which supply the intervals $\mathcal{C}(x)$ for titling. Neither the saliency annotations nor the creator chapter titles serve as references. The two tracks use separate videos but draw on the same set of personas.

We derive 24 personas from NVIDIA's Nemotron-Personas-USA dataset (Meyer & Corneil, 2025), a collection of synthetic personas. Each persona p is given as a persona card describing the viewer's background, interests, tastes, and inferred viewing preferences, such as topics to seek or avoid. The viewing preferences are based on the persona descriptions, not on observed viewing behavior. The 24 personas are not intended to represent the distribution of preferences in a viewer population.

We cluster personas by their taste descriptions into archetypes and take three personas from each of eight archetypes. We select the eight archetypes in an initial study that also generated chapter boundaries, favoring archetypes that produced different boundaries. Within each archetype, we take the most representative persona together with one close to it and one distant from it. Appendix A.2 details the card fields and selection procedures.

4.2 Reference Construction

We construct the reference annotations using two annotation models: gpt-oss-120b (Agarwal et al., 2025) (a_A) and Gemma-4-31B (Team et al., 2026) (a_B).

Highlights. Each clip x is divided into two-second units $\mathcal{U}(x)$, and a_B captions a representative frame of each unit without the persona. These captions are the only content the highlight annotation model sees. For each clip-persona pair, annotation model a_A runs three independent repetitions, each generating a reference perspective z^{ref} and then selecting units given the persona and z^{ref}. The perspective is omitted when the model rates the persona's interest in the clip as low, whereas an

Table 1: Comparison with representative video benchmarks. Our counts include only pairs with complete synthetic annotations.

Dataset	Scale	Conditioning	Output	Annotation
QVHighlights (Lei et al., 2021)	10,148 clips, 10,310 queries	Text query	Moments and saliency	Crowd
PHD-GIFs (García del Molino & Gygli, 2018)	119,938 videos, 222,015 GIFs	GIF history of 13,822 users	Personalized highlights	User GIFs
HIPPO-Video (Lee et al., 2025)	2,040 histories, 20,400 videos[a]	Simulated watch history	Personalized saliency	LLM simulation
VidChapters-7M (Yang et al., 2023a)	≈817K videos, ≈7M chapters	None	Chapter boundaries and titles	Creators
Ours: Highlights	165 source videos, 1,386 pairs	24 persona cards	Unit scores or selected units	LLM
Ours: Chapter titles	94 source videos, 315 pairs	17 persona cards	Titles on supplied intervals	LLM

[a]Saliency is annotated only on the last video of each history.

invalid perspective blocks selection. Each selection assigns units to two disjoint tiers, must-watch (MUST) and worth-watching (WORTH), and may be empty.

Let $\mathcal{S}^{(r)}_{\text{must}}$ and $\mathcal{S}^{(r)}_{\text{worth}}$ denote these tiers in repetition r, and let $\text{maj}(A_1, A_2, A_3)$ denote the unit indices contained in at least two of the sets A_r. The base and expanded references are as follows:

$$H_p^{\text{base}} = \text{maj}\big(\mathcal{S}^{(1)}_{\text{must}}, \mathcal{S}^{(2)}_{\text{must}}, \mathcal{S}^{(3)}_{\text{must}}\big), \tag{5}$$

$$H_p^{\text{exp}} = \text{maj}\big(\mathcal{S}^{(1)}_{\text{must}} \cup \mathcal{S}^{(1)}_{\text{worth}}, \ldots, \mathcal{S}^{(3)}_{\text{must}} \cup \mathcal{S}^{(3)}_{\text{worth}}\big). \tag{6}$$

References require all three repetitions to pass structural checks, and valid empty selections remain eligible, whereas failures never become empty references. Model evaluation uses these aggregated sets, while Section 4.4 also compares individual repetitions.

Chapter titles. For titling, $\mathcal{C}(x)$ contains the creator chapter intervals that have positive duration and at least one overlapping ASR utterance. For interval c_k, let E_k denote the ASR utterances overlapping c_k, with their identifiers and timestamps. Two annotation models a_A and a_B independently receive E_k and the persona p:

$$\left(t^{(j)}_{p,k}, \mathcal{E}^{(j)}_{p,k}\right) = a_j(E_k, p), \qquad j \in \{A, B\}, \tag{7}$$

where $t^{(j)}_{p,k}$ is a title and $\mathcal{E}^{(j)}_{p,k} \subseteq E_k$ contains the utterances cited as its support. Neither model receives a perspective, a creator title, or the other model's output, so the title references rely on the spoken content of each interval.

We keep the two titles as separate references rather than merging them. Let $v^{(j)}_{p,k} \in \{0, 1\}$ indicate whether a response contains a nonempty title and passes the evidence and schema checks. The available references are as below:

$$\mathcal{R}_{p,k} = \left\{ \left(j, t^{(j)}_{p,k}\right) \;\middle|\; j \in \{A, B\}, \quad v^{(j)}_{p,k} = 1 \right\}. \tag{8}$$

An annotation model may abstain when the evidence cannot support a title. Abstentions and invalid responses are not retried and leave the reference unavailable.

In both tasks, passing structural checks does not establish semantic correctness. Appendix B provides the prompts and construction settings.

4.3 Dataset Composition and Splits

The input collection pairs personas with 173 highlight clips and 195 chapter-titling videos, giving 1,480 and 775 video-persona pairs. Pairs are retained from an earlier interest screen, so each persona is paired only with videos rated as relevant to its interests, and titling covers 17 of the 24 personas. Complete references, which require all three highlight repetitions or valid titles from both annotation models for every interval, cover 1,386 highlight pairs from 165 source videos and 315 title pairs from 94 source videos. Each track is split by source video with personas shared across splits. The two tracks share no videos, and models are evaluated on the test split. Appendix A.3 reports per-split statistics and why title pairs are complete less often. Table 1 compares the benchmark with representative video benchmarks.

Table 2: Synthetic annotation diagnostics on the test split. (a) Persona contrast Δ_{persona} within and across annotation models. (b) Contrast under Direct (D) and Perspective (P) generation and its change. Brackets give 95% intervals, and † marks the only change that excludes zero after adjustment for the four comparisons.

(a) Persona contrast

	Highlights (F1 score)		Chapter titles	
Comparison	Base	Expanded	CIDEr-D	METEOR
Within a_A (GPT-OSS)	.195 [.136, .267]	.155 [.104, .215]	1.006 [.728, 1.290]	.083 [.061, .108]
Within a_B (Gemma)	.297 [.172, .429]	.474 [.350, .593]	2.010 [1.445, 2.598]	.216 [.159, .274]
Across a_A and a_B	.072 [.026, .125]	.073 [.024, .124]	.007 [−.031, .052]	.003 [.001, .006]

(b) Direct versus perspective-conditioned annotation

	Highlights: F1 score (base)			Chapter titles: METEOR		
Annotation model	D	P	Change	D	P	Change
a_A (GPT-OSS)	.192	.219	+.028 [−.060, .109]	.096	.126	+.029 [−.002, .058]
a_B (Gemma)	.316	.433	+.117† [.041, .215]	.199	.244	+.045 [−.001, .091]

4.4 Synthetic Annotation Diagnostics

We test whether annotations generated for the same persona agree more closely than annotations generated for different personas on the same content. Let $\bar{\sigma}_{\text{same}}$ denote the mean similarity between two annotations generated for the same persona, and $\bar{\sigma}_{\text{other}}$ the mean similarity between annotations for the target persona and for a fixed other persona. The persona contrast is as follows:

$$\Delta_{\text{persona}} = \bar{\sigma}_{\text{same}} - \bar{\sigma}_{\text{other}}. \tag{9}$$

We measure similarity between two synthetic annotations using F1 score between their selected highlight sets, separately for the base and expanded sets, and CIDEr-D (Vedantam et al., 2015) and METEOR (Banerjee & Lavie, 2005) between their titles. Appendix D details the metric implementations. Within-model comparisons use the three highlight repetitions of a_A, together with additional test-split generations that follow the construction protocol for a_B highlights and for titles from both models. All comparisons in Table 2(a) within a task share one set of available test targets, namely 143 of 229 highlight pairs from 22 source videos and 556 of 596 title targets from 21 videos, and confidence intervals resample source videos.

Persona contrast. Table 2(a) shows positive contrasts within each annotation model for both tasks, from 0.20 to 0.30 in base highlight F1 score and from 0.08 to 0.22 in title METEOR, all with intervals above zero. Across the two models, the highlight contrast falls to 0.07 and the title contrast nearly vanishes, at 0.003 in METEOR. Persona-associated differences repeat within each model, but the two models share little persona-specific title wording. Because F1 score equals one when both selections are empty, part of the highlight agreement reflects shared empty selections rather than shared selected units.

Effect of perspectives. Table 2(b) compares two fresh generation conditions on the test split: Direct, which uses the persona and supplied video evidence without a perspective, and Perspective, which additionally uses a perspective generated as in Section 3.3. It reports the persona contrast under each condition and its change for base highlights and title METEOR, using availability masks that differ from those of panel (a). All four changes have positive point estimates, but only the base-highlight change for a_B, marked †, has an adjusted 98.75% confidence interval entirely above zero. Perspectives lower both same- and other-persona agreement, the latter more strongly, and Appendix C shows positive point estimates for within-model title-contrast changes in CIDEr-D and ROUGE-L, no consistent change for expanded highlight sets, and more empty selections under Direct generation. The point estimates suggest larger persona contrasts with perspectives, but these increases do not by themselves establish improved annotation quality. These diagnostic generations do not alter the released references.

Table 3: Highlight selection against the references of a_A. None denotes persona-free outputs, and brackets give 95% intervals. †Trained on QVHighlights, which covers 17 of our 27 test source videos. ‡Most outputs fail.

	None		Target persona		Target − other persona	
Model	Base	Exp.	Base	Exp.	Base	Expanded
Score outputs (AP)						
CLIP	–	–	.151	.230	+.017 [.006, .029]	+.017 [.005, .031]
SigLIP2	–	–	.115	.183	+.020 [.008, .034]	+.014 [.001, .028]
QD-DETR†	–	–	.231	.299	+.032 [.003, .066]	+.015 [−.005, .038]
CG-DETR†	–	–	.198	.268	+.035 [.015, .061]	+.031 [.013, .053]
Set outputs (F1 score)						
TimeChat	.090	.167	.085	.175	−.005 [−.016, .005]	+.004 [−.007, .016]
Qwen3-VL‡	.092	.173	.049	.070	−.004 [−.014, .004]	−.015 [−.028, −.002]

5 Experiments

5.1 Experimental Setup

Baselines. We evaluate existing models without benchmark-specific training, spanning image–text embedding models, query-conditioned highlight detectors, video language models, and text language models. For highlights, CLIP (Radford et al., 2021) and SigLIP2 (Tschannen et al., 2025) score each unit by the similarity between its representative frame and the interest entries of the persona card, which are encoded separately to fit the text encoder. QD-DETR (Moon et al., 2023b) and CG-DETR (Moon et al., 2023a), trained on QVHighlights, use each interest entry as a text query with their CLIP-feature checkpoints and average the resulting scores. TimeChat (Ren et al., 2024) and Qwen3-VL (Bai et al., 2025) generate time spans from sampled frames and the persona card, and we map each span to the units it overlaps.

For titles, Llama-3.1 (Grattafiori et al., 2024), Mistral-Nemo (Mistral AI & NVIDIA, 2024), and Qwen3 (Yang et al., 2025) generate one title per interval from its overlapping ASR and the persona card. Chapter-Llama (Ventura et al., 2025) receives the same inputs through a fixed prompt adapted to supplied intervals, and Vid2Seq (Yang et al., 2023b) generates persona-free titles from frames and ASR given the interval boundaries. Appendix D describes these adaptations, input limits such as ASR truncation in Vid2Seq, and the overlap between baseline training data and our test videos.

Controls. Each persona-conditioned baseline receives the target persona, whose references are used for scoring. To test whether this persona matters, we score against the same references the output produced for another persona paired with the same content, chosen by a fixed cyclic shift, and pairs without such a persona are excluded from this contrast. Generative baselines also produce one persona-free output per clip or interval, which is scored against the references of every target persona to test the effect of adding a persona at all. Prediction failures remain in the denominators rather than being excluded.

Evaluation. For highlights, a unit counts as relevant if it belongs to H_p^{base}, or to H_p^{exp} for the expanded variant. We compute AP for score outputs and F1 score for set outputs, and AP excludes pairs whose reference is empty. Titles are scored with CIDEr-D (Vedantam et al., 2015) and METEOR (Banerjee & Lavie, 2005) separately against the references of a_A and a_B on complete pairs. Scores are averaged over personas within clips, clips within source videos, and source videos with equal weight, and confidence intervals resample source videos.

5.2 Highlight Selection

Table 3 evaluates score outputs with AP and set outputs with F1 score against the references of a_A. All four score-output models score higher with the target persona than with another persona, with base contrasts of 0.017 to 0.035 whose intervals exclude zero. These gains are small relative to base AP of 0.115 to 0.231. QD-DETR and CG-DETR achieve the highest base and expanded AP, but their

Table 4: Supplied-interval chapter titling in METEOR against the references of a_A and a_B. None denotes persona-free outputs, brackets give 95% intervals, and CIDEr-D appears in Appendix E. ‡Adapted to supplied intervals. §Persona-free, with partly truncated ASR.

	a_A references			a_B references		
Model	None	Target	Target − other	None	Target	Target − other
Llama-3.1	.131	.116	+.003 [−.007, .014]	.232	.178	.000 [−.015, .018]
Mistral-Nemo	.145	.144	+.011 [.005, .017]	.248	.211	−.004 [−.011, .002]
Qwen3	.144	.158	+.007 [−.001, .016]	.248	.232	+.003 [−.005, .012]
Chapter-Llama‡	.063	.081	−.005 [−.017, .002]	.116	.138	−.004 [−.022, .008]
Vid2Seq§	.060	–	–	.121	–	–

QVHighlights training data include 17 of our 27 test source videos, so part of their advantage may reflect exposure.

Neither set-output model shows a consistent gain over another persona. TimeChat shows no reliable contrast, and Qwen3-VL scores lower with a persona than without one. This gap coincides with failures, which affect 147 of 252 persona-conditioned Qwen3-VL outputs but only 8 of 28 persona-free outputs, mostly through the output-length limit.

Consistent gains over another persona appear in base AP for the score-output models but not in F1 for the set-output models, although the two groups also differ in architecture, input, and metric.

5.3 Supplied-Interval Chapter Titling

Table 4 scores titles for fixed supplied intervals separately against the references of a_A and a_B. Adding a persona does not consistently raise agreement. The three general language models mostly score lower with the target persona than without one, most clearly against the references of a_B, and only Qwen3 improves, against the references of a_A alone. Chapter-Llama shows the opposite pattern, but its persona-free outputs also fail more often, 99 of 225 versus 188 of 787 persona-conditioned outputs.

Target-minus-other contrasts are close to zero for most models and references. The one clear positive contrast, Mistral-Nemo against the references of a_A, becomes negative against those of a_B. This dependence on the annotation model is consistent with the near-zero cross-model title contrast in Section 4.4. Chapter-Llama is applied to a setting it was not trained for, and Vid2Seq receives no persona and partly truncated ASR, so their lower agreement cannot be separated from these constraints.

6 Limitations

Our references are synthetic annotations without human validation, so agreement with them measures agreement with model-generated, persona-conditioned references rather than factual correctness or the preferences of real viewers. Because this consistency can depend on the annotation model, we keep the references of a_A and a_B separate and report where their results diverge. The benchmark uses a controlled inventory of 24 synthetic personas and fixed title intervals, which isolates persona effects on the same content but does not cover viewer populations or persona-specific boundaries. Testing whether these patterns hold under human judgments and for broader content and viewers remains future work.

7 Conclusion

We introduce a benchmark for persona-conditioned highlight selection and chapter titling that compares personas on the same video content. Persona-associated differences in its synthetic annotations repeat within an annotation model, but the two annotation models share little persona-specific title wording. Among existing models, highlight models that score every unit show consistent gains over another persona in base AP, and titling effects depend on the annotation model used for the references.

AI USE STATEMENT

We used gpt-oss-120b and Gemma-4-31B to generate the synthetic reference annotations, unit captions, and diagnostic generations described in Section 4.2 and Appendix B. The persona cards include viewing preferences inferred with a language model from the source persona descriptions. We also used LLM assistants to draft and edit parts of the manuscript for clarity. The authors reviewed all AI-assisted content and take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

Our personas are synthetic profiles derived from Nemotron-Personas-USA, released under CC BY 4.0, and do not describe real individuals. Because they include demographic attributes, persona-conditioned outputs may reflect stereotypes, so we group personas by stated interests rather than demographics and avoid demographic labels in presenting archetypes. The personas are not a representative sample of any population, and agreement with their annotations should not be read as the preferences of real groups. No human subjects were involved. The source videos come from publicly released datasets, and we release annotations and identifiers rather than video files, following the terms of those datasets. Persona-conditioned personalization could also narrow what viewers see, so deployed systems should let viewers inspect and change the conditioning information.

REPRODUCIBILITY STATEMENT

Section 3 defines the tasks, and Sections 4.2 and 4.3 with Appendices A and B describe the data, personas, verbatim prompts, generation settings, and request outcomes. Section 5.1 and Appendix D list baseline checkpoints, inputs, decoding settings, scoring rules, and the bootstrap procedure with its seed. Because language-model generation may not be exactly repeatable, we release the generated annotations, persona cards, and perspectives themselves rather than only the procedure to regenerate them. All reported tables are regenerated from saved outputs by scripts that check their values against the reported results.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 82–92, 2021.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8107–8117. IEEE, 2021.

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.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Xiao Cao, Zitan Chen, Canyu Le, and Lei Meng. Multi-modal video chapter generation. *arXiv preprint arXiv:2209.12694*, 2022.

Runnan Chen, Penghao Zhou, Wenzhe Wang, Nenglun Chen, Pai Peng, Xing Sun, and Wenping Wang. PR-Net: Preference reasoning for personalized video highlight detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7960–7969. IEEE, 2021.

Ana García del Molino and Michael Gygli. PHD-GIFs: Personalized highlight detection for automatic GIF creation. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 600–608, 2018.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

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

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.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, 2024.

Jeongeun Lee, Youngjae Yu, and Dongha Lee. HIPPO-Video: Simulating watch histories with large language models for history-driven video highlighting. In *Second Conference on Language Modeling*, 2025.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Yev Meyer and Dane Corneil. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions, June 2025. URL `https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA`.

Mistral AI and NVIDIA. Mistral NeMo. `https://mistral.ai/news/mistral-nemo/`, 2024. Accessed: 2026-09-25.

WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration for video temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023a.

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, 2023b.

Junfu Pu, Teng Wang, Yixiao Ge, Yuying Ge, Chen Li, and Ying Shan. ARC-Chapter: Structuring hour-long videos into navigable chapters and hierarchical summaries. *arXiv preprint arXiv:2511.14349*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Mrigank Rochan, Mahesh Kumar Krishna Reddy, Linwei Ye, and Yang Wang. Adaptive video highlight detection by learning from user history. In *European conference on computer vision*, pp. 261–278. Springer, 2020.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wührl, et al. Which demographics do LLMs default to during annotation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17331–17348, 2025.

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.

Jinhwan Sul, Jihoon Han, and Joonseok Lee. Mr. HiSum: A large-scale dataset for video highlight detection and summarization. *Advances in Neural Information Processing Systems*, 36:40542–40555, 2023.

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.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Chapter-Llama: Efficient chaptering in hour-long videos with LLMs. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18947–18958. IEEE, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. VidChapters-7M: Video chapters at scale. *Advances in Neural Information Processing Systems*, 36:49428–49444, 2023a.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

A Dataset Details

A.1 Video Sources and Preprocessing

Highlight clips come from QVHighlights (Lei et al., 2021). We define clip-local time from decoded video timestamps and divide each clip into two-second units up to the earlier of 150 seconds and the decoded video end, retaining a shorter final unit. Annotation model a_B captions a representative frame of each unit without a persona. Chapter-titling videos come from the `sml300_val` subset distributed with Chapter-Llama (Ventura et al., 2025), whose name refers to that source subset rather than to our validation split. We keep creator chapter intervals with positive duration and at least one overlapping ASR utterance, which excludes 11 intervals. Figure 3 summarizes the resulting clips and intervals.

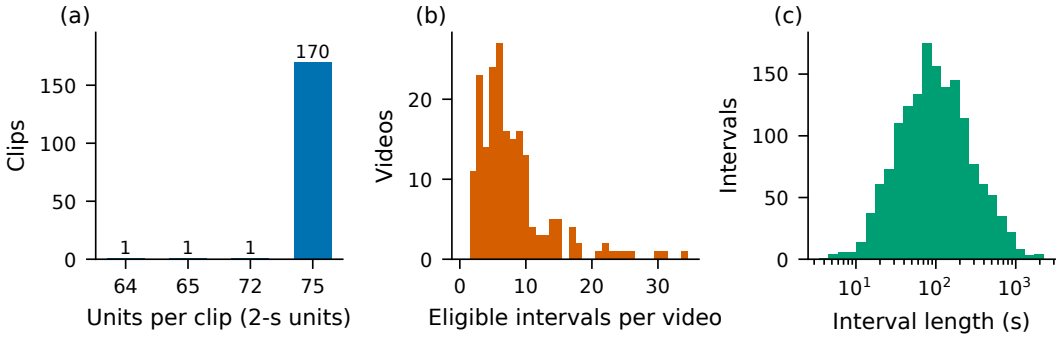


Figure 3: Video composition. (a) Units per highlight clip, with a median of 75 over 173 clips. (b) Eligible intervals per title video, with a median of 6 over 195 videos. (c) Title interval durations, with a median of 94 seconds over 1,561 intervals.

A.2 Persona Construction and Selection

We sample 80,000 adult personas from Nemotron-Personas-USA (Meyer & Corneil, 2025), stratified by age band and sex, embed their taste descriptions with `BAAI/bge-large-en-v1.5`, and cluster the embeddings with k-means into 20 clusters. After manual inspection, we remove two clusters that largely duplicate a third and merge two groups of closely related clusters, three into one and two into one, which leaves 15 candidate archetypes. An initial study generates chapter boundaries under each candidate archetype with repeated runs. Let F_{ij} denote the mean boundary F1 score between archetypes i and j, and F_{ii} the mean F1 score between repeated runs of archetype i. We measure the disagreement between two archetypes as $d_{ij} = F_{ij} - (F_{ii} + F_{jj})/2$, so that archetypes with unstable boundaries do not appear distinct because of run-to-run variation. Starting from the pair with the lowest d_{ij}, we greedily add the candidate with the lowest mean d_{ij} to the selected archetypes until eight are selected, and all 15 candidates remain eligible throughout. Within each archetype, we take three members by cosine similarity in the embedding space, without any model calls: the member closest to the cluster centroid, which we call the medoid, the member most similar to the medoid, and the member farthest from the centroid. Table 5 lists the personas, Listing 1 shows one persona card, and Figure 4 shows the pairs each persona receives.

Table 5: The 24 personas. Representative interests are the first two interest entries of each card, and pair counts cover all splits. The seven personas without title pairs are all archetype medoids.

Persona ID	Role	Age	Sex	Occupation	Region	#Int.	Representative interests	HL compl.	HL fixed	Title compl.	Title fixed
arch00-rep01	medoid	40	female	transportation worker	West	6	Gardening, Hiking local trails	81	86	0	0
arch00-rep02	near	59	female	secondary school teacher	Midwest	10	Reading historical fiction, Gardening	63	68	19	38
arch00-rep07	diverse	28	female	veterinary technologist or technician	South	10	Reading contemporary fiction, Gardening	66	71	18	46
arch01-rep01	medoid	35	male	sales representative of services	South	8	Visiting art galleries, Photography	99	107	0	0
arch01-rep02	near	21	male	not in workforce	South	8	Sketching, Visiting Art Galleries and Museums	88	97	40	91
arch01-rep07	diverse	20	male	first line supervisor of retail sales worker	Midwest	10	Digital illustration, Graphic design	58	64	42	93
arch02-rep01	medoid	45	female	janitor or building cleaner	Northeast	8	Visiting art galleries and museums, Watercolor painting	66	71	0	0
arch02-rep02	near	29	female	teaching assistant	South	10	Visiting Art Museums, Watercolor Painting	89	95	23	52
arch02-rep07	diverse	19	female	not in workforce	South	10	Painting, Sketching	57	64	22	52
arch03-rep01	medoid	44	male	business operations specialist	West	10	Hiking, Watching and playing soccer	50	50	0	0
arch03-rep02	near	44	male	elementary or middle school teacher	South	9	Playing soccer, Gardening	48	50	11	37
arch03-rep07	diverse	28	male	bartender	West	8	home mixology experimentation, reading cocktail history	55	62	21	47
arch05-rep01	medoid	42	male	construction laborer	Midwest	10	Fishing, Hunting	37	40	0	0
arch05-rep02	near	28	male	not in workforce	Northeast	6	Fishing, Hunting (seasonal)	39	43	16	39
arch05-rep07	diverse	49	male	automotive service technician or mechanic	Midwest	10	Classic car restoration, Scale model building	20	20	16	41
arch09-rep01	medoid	84	male	not in workforce	Northeast	8	Vegetable and flower gardening, Woodworking (birdhouses, furniture repair)	41	46	0	0
arch09-rep02	near	76	male	not in workforce	Midwest	9	Gardening (vegetables and roses), Woodworking (small furniture and birdhouses)	47	48	18	44
arch09-rep07	diverse	52	female	bartender	West	10	Local trivia nights, Community board-game gatherings	35	35	12	35
arch11-rep01	medoid	25	male	no occupation	Northeast	9	Attending local sports events, Pickup basketball	68	70	1	1
arch11-rep02	near	45	male	cashier	South	8	Attending local sports events, Listening to country and classic rock music	47	48	22	49
arch11-rep07	diverse	45	male	wholesale or retail buyer	West	6	Chess, Reading nonfiction	50	56	11	36
arch14-rep01	medoid	52	male	education or childcare administrator	Northeast	8	Gardening, Chess	76	80	0	0
arch14-rep02	near	52	male	software developer	Northeast	10	Gardening, Reading nonfiction (technology, history, finance)	65	67	18	53
arch14-rep07	diverse	70	male	not in workforce	Midwest	7	daily vegetable gardening, playing bridge	41	42	5	21

Listing 1: Persona card `arch00-rep01`, with at most three viewing preferences to seek and to avoid.

```
Persona card: arch00-rep01 (card_version v1, role medoid)

DEMOGRAPHICS
  age: 40
  sex: female
  occupation: transportation worker
  region: West

INTERESTS
  - Gardening
  - Hiking local trails
  - Reading historical fiction
  - Cooking seasonal meals
  - Participating in a small book club
  - Volunteering at the neighborhood food pantry

TASTE_SKETCH
  sports: low-intensity outdoor activities; hiking foothill trails Hiking Club; community yoga
        classes; solitary walks over team sports; smartwatch to track steps
  arts: historical fiction; discussing works Hilary Mantel; Philippa quiet book club; Art
       Museum's exhibitions; solace listening folk singer Sufjan; seasonal meals
  culinary: cooks with a farm-to-table philosophy; turning garden harvests dishes; goat-cheese
        salad; thyme; wild mushroom risotto flavored rosemary; hearty sourdough loaf
  travel: road trips balance adventure family; weekend escapes; summer vacation camping; cross
       -country drive; budgeting travel expenses dedicated spreadsheet

VIEWING_INFERENCE (seek/avoid: first 3 entries each, verbatim)
  seek:
    - organic gardening tutorials -- matches gardening interest and farm-to-table cooking
         philosophy
    - foraging wild mushroom guides -- aligns with love of seasonal meals and mushroom risotto
    - historical fiction book club discussions -- directly reflects reading historical fiction
          and small book club participation
  avoid:
    - high-calorie BBQ grill competitions -- conflicts with farm-to-table, seasonal, health-
         focused cooking preferences

[viewing_inference truncated: seek 3/8, avoid 1/1; generated by openai/gpt-oss-120b,
     prompt_version v3]
```

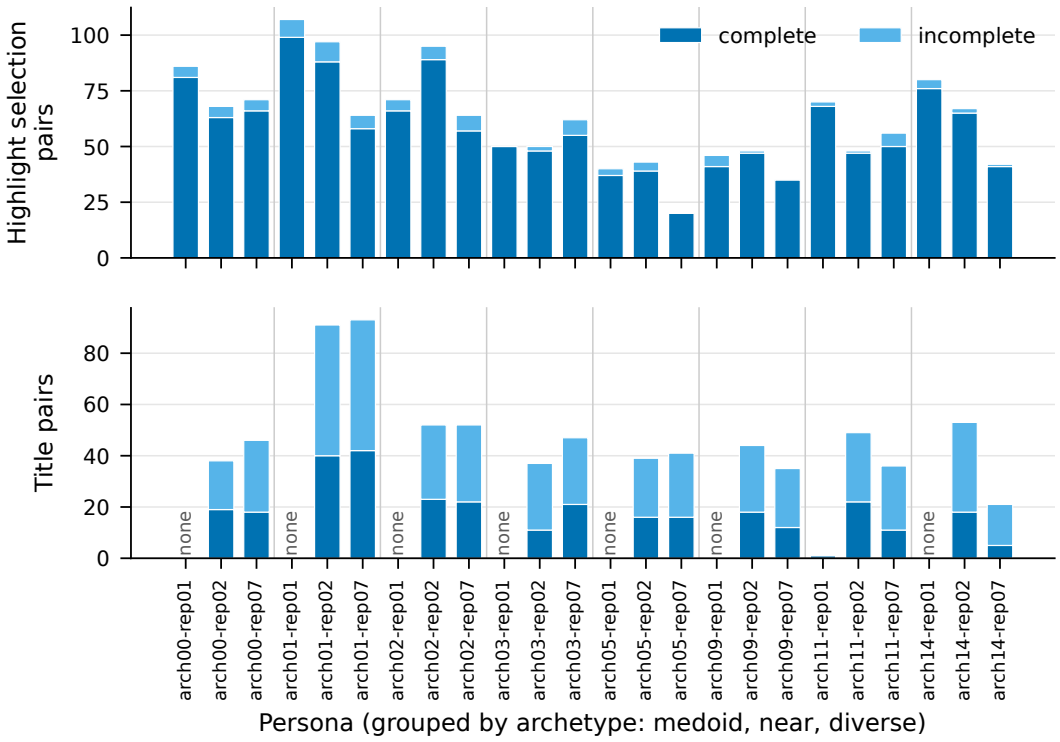


Figure 4: Complete and incomplete pairs per persona. Vertical rules separate archetypes, and seven personas have no title pairs.

A.3 Detailed Dataset Statistics

Table 6 separates the fixed input collection from pairs with complete synthetic annotations. A pair combines a clip with a persona for highlights, or a video with a persona for chapter titles. Both collections retain pairs that passed an earlier interest screen. Splits are disjoint by source video within each track, and personas are shared across splits. Incomplete pairs keep their recorded generation outcomes and are neither replaced nor converted into empty annotations. A title pair is complete only when both annotation models provide valid titles for all of its intervals, so title completion is lower even though 11,365 of 12,586 interval-level title references are available (Figure 5).

Table 6: Split statistics. A source video counts as complete when it has at least one complete pair, and completeness indicates structural availability, not human-verified quality.

		Source videos		Pairs	
Task	Split	Input	Complete	Input	Complete
Highlight selection	train	116	113	1,017	944
Highlight selection	val	25	25	211	200
Highlight selection	test	27	27	252	242
Highlight selection	Total	168	165	1,480	1,386
Chapter titling	train	136	62	511	200
Chapter titling	val	29	16	144	62
Chapter titling	test	30	16	120	53
Chapter titling	Total	195	94	775	315

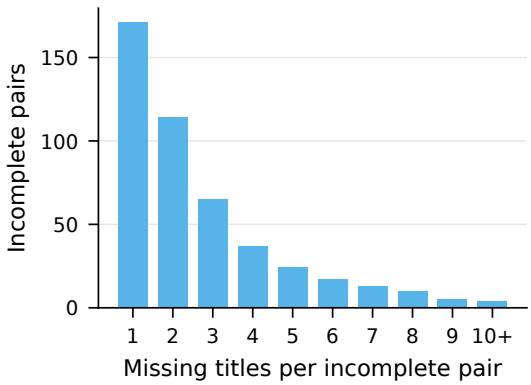


Figure 5: Missing interval-level titles per incomplete title pair. Of 460 incomplete pairs, 171 miss a single title.

B REFERENCE CONSTRUCTION

Table 7 lists the settings of each generation stage, and Table 8 reports the outcome of every request. Structural checks verify unit or utterance coverage, evidence identifiers, and response schemas, and they do not assess semantic correctness. Figure 6 summarizes the size of the resulting references, and the prompt templates follow verbatim with braces marking inserted inputs.

Table 7: Generation settings. Rows marked Diag. produce only the diagnostic annotations of Appendix C.

Stage	Model	Input	Output	Temp.	Top-p	Input cap	Max new tokens	Seed(s)	Repetitions (requests)
Unit caption	a_B (Gemma-4-31B-it)	one representative frame per 2-s unit + fixed instruction	free text	0	default	4,096	256	2026091901	1 per unit (12,951)
Perspective	a_A (gpt-oss-120b, reasoning low)	persona card + all unit captions	JSON {interest 0-3, z}	0.3	default	16,000	400	2026091901, 2026091902, 2026091903	3 pipelines per pair (4,440)
Highlight selection	a_A (gpt-oss-120b, reasoning low)	persona card + same-pipeline perspective + all unit captions	JSON {must, worth, evidence}	0.3	default	16,000	1,200	2026091901, 2026091902, 2026091903	3 pipelines per pair (4,440), unit kept if $\geq$2 of 3
Title	a_A and a_B	persona card + all ASR utterances overlapping the interval	JSON {status, title, evidence IDs}	0.3	default	24,576	512	2026091901	1 per model per pair-interval (6,293 + 6,293)
Diag.: perspective + HL selection	a_B	same as reference HL stages (same prompts)	same as reference	0.3	default	16,000	400 / 1,200	per-request seed	3 pipelines per test target (714 + 714)
Diag.: title repeat	a_A and a_B	identical title prompt, new seed	same as reference	0.3	default	24,576	512	per-request seed	1 extra draw (615 + 615)
Diag.: Direct / Perspective	a_A and a_B	D: persona + evidence, P: persona + generated perspective z + evidence	JSON (HL selection or title)	0.3	default	context limit	z 768, HL 4,096, title 1,024	D and P final share a seed, z distinct	HL 1,512 D + 1,512 z + 1,512 P, title 3,148 D + 3,148 z + 3,148 P

Table 8: Request outcomes for the released references. Valid denotes schema validity only. Null (low interest) counts perspectives withheld because the persona's rated interest is low, and Input verified counts requests whose input is reconstructed exactly from the released inputs.

Stage	Model	Requests	Valid	Valid empty	Null (low interest)	Abstention	Malformed	Output limit	Blocked	Input verified
Unit caption	a_B	12,951	12,948	0	0	0	0	3	0	12,944/12,944
Perspective	a_A	4,440	2,794	0	1,570	0	1	0	75	4,365/4,365
Highlight selection	a_A	4,440	3,725	566	0	0	33	40	76	4,364/4,364
Title	a_A	6,293	5,195	0	0	692	0	406	0	6,293/6,293
Title	a_B	6,293	6,170	0	0	38	46	39	0	6,293/6,293

Listing 2: Unit captioning prompt, sent to a_B with the unit frame.

```
Describe only visible facts in this single frame. Do not infer unseen events, intentions, or
    persona preferences. Be concise.
```

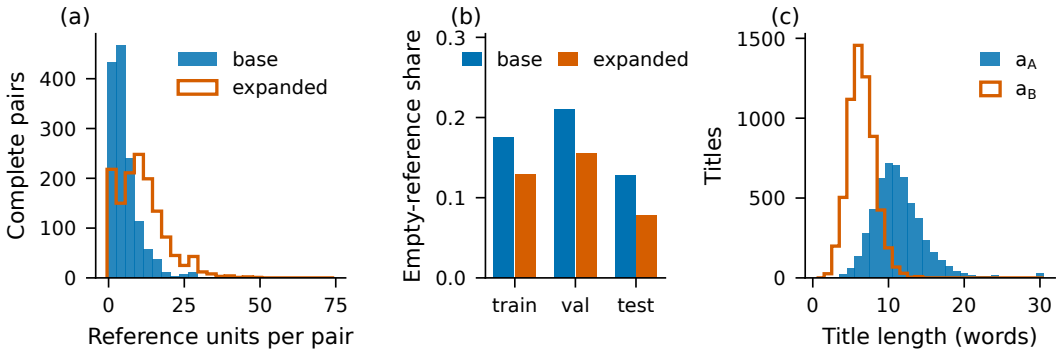


Figure 6: Reference size. (a) Units in complete highlight references, with medians of 4 for base and 10 for expanded. (b) Share of empty highlight references by split. (c) Title length in words, with medians of 11 for a_A and 6 for a_B.

Listing 3: Reference perspective prompt.

```
Assess this target persona's relevance to the supplied video evidence. Do not invent facts or
    force differences between people.
PERSONA
{PERSONA}
CAPTION UNITS
{CAPTION_UNITS}
Return JSON {"interest":0..3,"z":string or null}, If interest<=1 return null, otherwise a
    grounded auxiliary perspective. No title or annotation from another generator is
    supplied.
```

Listing 4: Highlight selection prompt. A low-interest perspective is replaced by a fixed instruction to apply the same persona.

```
Select moments relevant to the target persona from ALL supplied caption units. Persona
    preferences never establish events absent from evidence. Do not force different
    selections for different people. Empty selections are allowed if unsupported. Do not
    select for a different audience.
PERSONA
{PERSONA}
AUXILIARY PERSPECTIVE
{AUXILIARY_PERSPECTIVE}
CAPTION UNITS
{CAPTION_UNITS}
Return JSON {"must": [unit IDs], "worth": [unit IDs], "evidence": {selected_unit_id:[
    supporting supplied unit IDs]}}. MUST and worth are disjoint; every selected unit needs
    evidence including its own unit ID. Schema validity is not semantic verification.
```

Listing 5: Chapter title prompt.

```
Write one chapter title for this target persona and this fixed interval using ALL provided ASR
     utterances. ASR is speech evidence, not proof of unseen visual events. Persona is not a
    source of video facts. Boundary-crossing utterances are marked; word-level timestamps are
     unavailable. Do not force persona-specific differences. Do not use creator titles or
    another generator's title.
PERSONA
{PERSONA}
FULL INTERVAL ASR
{FULL_INTERVAL_ASR}
Return JSON {"status":"supported" or "insufficient_evidence", "title":string or null, "
    evidence_utterance_ids":[provided utterance IDs]}. If evidence cannot support a title,
    return insufficient_evidence with title null. This is not a missing/malformed response.
```

C Synthetic Annotation Diagnostics

Within an annotation model, same-persona agreement compares two independent generations for a persona, and other-persona agreement compares aligned generations for the target persona and its cyclic other persona. Across models, agreement averages both directions. Table 9 reports the values behind Table 2(a), Table 10 reports every Direct–Perspective comparison, and Table 11 shows that Direct highlight annotations contain more empty selections. Table 12 gives the corresponding results on the original, larger populations, and Figure 7 shows the spread across source videos. The Direct–Perspective prompts differ from the reference-construction templates, and this analysis was not preregistered.

Table 9: Same- and other-persona agreement behind Table 2(a), with 95% intervals from 2,000 source resamples.

Comparison	Task	Metric	Same	Other	Δ	95% CI	N (scorable/planned)	Sources
within a_A	HL	F1 base	0.5871	0.3921	+0.1950	[0.1356, 0.2672]	143/229 pairs	22
within a_A	HL	F1 expanded	0.6181	0.4628	+0.1552	[0.1043, 0.2149]	143/229 pairs	22
within a_A	Title	CIDEr-D	3.6647	2.6590	+1.0057	[0.7279, 1.2900]	556/596 targets	21
within a_A	Title	METEOR	0.3887	0.3053	+0.0834	[0.0606, 0.1077]	556/596 targets	21
within a_A	Title	ROUGE-L	0.5572	0.4683	+0.0890	[0.0649, 0.1128]	556/596 targets	21
within a_B	HL	F1 base	0.8466	0.5493	+0.2973	[0.1717, 0.4292]	143/229 pairs	22
within a_B	HL	F1 expanded	0.9089	0.4350	+0.4739	[0.3500, 0.5926]	143/229 pairs	22
within a_B	Title	CIDEr-D	8.5004	6.4906	+2.0098	[1.4446, 2.5982]	556/596 targets	21
within a_B	Title	METEOR	0.8695	0.6531	+0.2163	[0.1586, 0.2744]	556/596 targets	21
within a_B	Title	ROUGE-L	0.9223	0.7941	+0.1282	[0.0886, 0.1683]	556/596 targets	21
across a_A/a_B	HL	F1 base	0.3020	0.2305	+0.0716	[0.0259, 0.1253]	143/229 pairs	22
across a_A/a_B	HL	F1 expanded	0.4281	0.3551	+0.0729	[0.0240, 0.1239]	143/229 pairs	22
across a_A/a_B	Title	CIDEr-D	1.2353	1.2285	+0.0068	[−0.0315, 0.0523]	556/596 targets	21
across a_A/a_B	Title	METEOR	0.2064	0.2033	+0.0031	[0.0008, 0.0059]	556/596 targets	21
across a_A/a_B	Title	ROUGE-L	0.3309	0.3268	+0.0041	[0.0004, 0.0082]	556/596 targets	21

Table 10: Direct–Perspective comparisons for all metrics. Adjusted 98.75% intervals apply to the four primary comparisons, and other rows carry unadjusted 95% intervals.

Task	Comparison	Metric	Same_D	Other_D	Δ_D	Same_P	Other_P	Δ_P	Change	95% CI	Primary	98.75% CI	N
HL	within a_A	F1 base	0.6248	0.4330	0.1919	0.5201	0.3007	0.2195	+0.0276	[−0.0605, 0.1090]	yes	[−0.0786, 0.1244]	247/249
HL	within a_A	F1 expanded	0.5922	0.4174	0.1748	0.5443	0.3653	0.1790	+0.0041	[−0.0705, 0.0918]	no	–	247/249
HL	within a_B	F1 base	0.9529	0.6368	0.3161	0.8862	0.4530	0.4332	+0.1171	[0.0412, 0.2154]	yes	[0.0256, 0.2495]	203/249
HL	within a_B	F1 expanded	0.9371	0.4719	0.4651	0.8897	0.4740	0.4157	−0.0495	[−0.1206, 0.0126]	no	–	203/249
HL	across a_A/a_B	F1 base	0.4895	0.3979	0.0916	0.2558	0.1635	0.0923	+0.0007	[−0.0438, 0.0503]	no	–	202/249
HL	across a_A/a_B	F1 expanded	0.4137	0.2939	0.1197	0.4214	0.2975	0.1240	+0.0042	[−0.0597, 0.0661]	no	–	202/249
Title	within a_A	CIDEr-D	5.0355	3.8937	1.1418	3.9516	2.5992	1.3524	+0.2106	[−0.0428, 0.4536]	no	–	481/730
Title	within a_A	METEOR	0.4834	0.3869	0.0965	0.4048	0.2792	0.1256	+0.0292	[−0.0017, 0.0585]	yes	[−0.0078, 0.0699]	481/730
Title	within a_A	ROUGE-L	0.6728	0.5795	0.0933	0.5783	0.4435	0.1348	+0.0415	[0.0122, 0.0697]	no	–	481/730
Title	within a_B	CIDEr-D	8.1153	6.0895	2.0258	7.2624	4.8730	2.3894	+0.3636	[−0.1633, 0.8910]	no	–	725/730
Title	within a_B	METEOR	0.8035	0.6047	0.1987	0.7195	0.4757	0.2438	+0.0451	[−0.0006, 0.0911]	yes	[−0.0132, 0.1091]	725/730
Title	within a_B	ROUGE-L	0.8930	0.7626	0.1304	0.8476	0.6741	0.1735	+0.0431	[0.0158, 0.0717]	no	–	725/730
Title	across a_A/a_B	CIDEr-D	2.0363	1.9624	0.0739	1.6161	1.5712	0.0449	−0.0289	[−0.1131, 0.0465]	no	–	480/730
Title	across a_A/a_B	METEOR	0.2543	0.2479	0.0064	0.2113	0.2044	0.0069	+0.0005	[−0.0077, 0.0069]	no	–	480/730
Title	across a_A/a_B	ROUGE-L	0.4113	0.4010	0.0103	0.3480	0.3393	0.0087	−0.0016	[−0.0147, 0.0097]	no	–	480/730

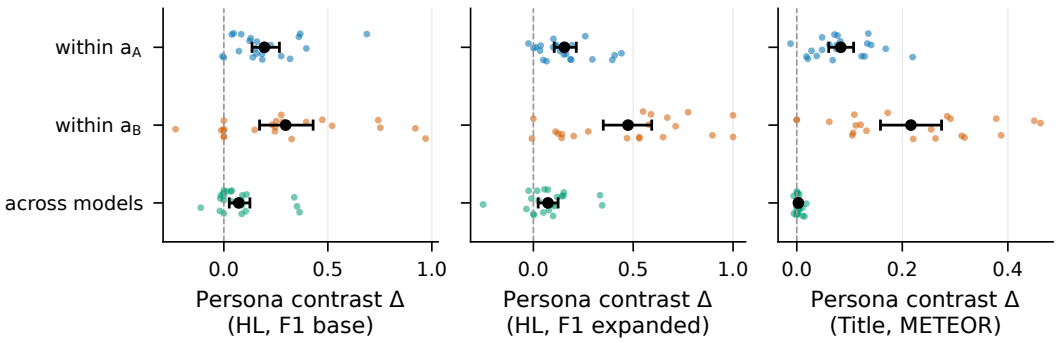


Figure 7: Persona contrast per source video. Dots are source videos, and black markers show the mean and 95% interval of Table 2(a).

Table 11: Completion and empty selections under Direct and Perspective generation. (a) Outcome rates per target. For a_A, valid empty highlight selections fall from 0.400 under Direct to 0.129 under Perspective. (b) Shares of annotation comparisons in which both, one, or neither selection is empty.

(a) Outcome rates

Task	Model	Outcome	D rate	P rate	P − D	95% CI	Targets
HL	a_A	supported (nonempty selection / title)	0.600	0.870	+0.2698	[0.1819, 0.3705]	252
HL	a_A	valid empty selection	0.400	0.129	−0.2709	[−0.3705, −0.1824]	252
HL	a_A	evidence mismatch	0.000	0.001	+0.0010	[0.0000, 0.0031]	252
HL	a_B	supported (nonempty selection / title)	0.800	0.872	+0.0714	[−0.0253, 0.1800]	252
HL	a_B	valid empty selection	0.169	0.064	−0.1052	[−0.1981, −0.0334]	252
HL	a_B	output limit	0.001	0.000	−0.0009	[−0.0026, 0.0000]	252
HL	a_B	blocked (perspective failed)	0.000	0.007	+0.0069	[0.0000, 0.0171]	252
HL	a_B	evidence mismatch	0.014	0.033	+0.0185	[−0.0165, 0.0741]	252
HL	a_B	out-of-range ID	0.004	0.003	−0.0010	[−0.0031, 0.0000]	252
HL	a_B	duplicate ID	0.000	0.003	+0.0026	[0.0000, 0.0079]	252
HL	a_B	overlapping selection classes	0.008	0.018	+0.0100	[−0.0016, 0.0245]	252
HL	a_B	parse error	0.003	0.001	−0.0023	[−0.0073, 0.0016]	252
Title	a_A	supported (nonempty selection / title)	0.907	0.925	+0.0183	[−0.0046, 0.0410]	787
Title	a_A	insufficient-evidence abstention	0.093	0.075	−0.0183	[−0.0410, 0.0046]	787
Title	a_B	supported (nonempty selection / title)	0.996	0.990	−0.0052	[−0.0146, 0.0019]	787
Title	a_B	insufficient-evidence abstention	0.003	0.003	0.0000	[0.0000, 0.0000]	787
Title	a_B	blocked (perspective failed)	0.000	0.007	+0.0066	[0.0000, 0.0160]	787
Title	a_B	parse error	0.001	0.000	−0.0014	[−0.0043, 0.0000]	787

(b) Empty selections in pairwise comparisons

			Direct (D)			Perspective (P)			
Model	Metric	Comparison	Both empty	One empty	Both nonempty	Both empty	One empty	Both nonempty	Comparisons
within a_A	F1 base	same persona	0.350	0.199	0.451	0.080	0.137	0.782	1,482
within a_A	F1 base	other persona	0.280	0.339	0.381	0.063	0.173	0.764	1,482
within a_A	F1 expanded	same persona	0.270	0.196	0.535	0.059	0.116	0.824	1,482
within a_A	F1 expanded	other persona	0.204	0.328	0.469	0.052	0.131	0.817	1,482
within a_B	F1 base	same persona	0.564	0.034	0.402	0.411	0.080	0.509	1,218
within a_B	F1 base	other persona	0.456	0.240	0.304	0.266	0.352	0.383	1,218
within a_B	F1 expanded	same persona	0.161	0.009	0.830	0.083	0.006	0.911	1,218
within a_B	F1 expanded	other persona	0.068	0.181	0.751	0.049	0.065	0.886	1,218
across a_A/a_B	F1 base	same persona	0.366	0.319	0.315	0.124	0.357	0.519	3,636
across a_A/a_B	F1 base	other persona	0.323	0.389	0.288	0.100	0.390	0.510	3,636
across a_A/a_B	F1 expanded	same persona	0.138	0.272	0.590	0.055	0.096	0.849	3,636
across a_A/a_B	F1 expanded	other persona	0.094	0.338	0.568	0.047	0.106	0.847	3,636

Table 12: Diagnostics on the original populations before restricting to targets available in every comparison.

Comparison	Metric	Same	Other	Δ	95% CI	N	Sources
within a_A (aligned draws)	F1 base	0.5294	0.3690	+0.1604	[0.1262, 0.1949]	229 targets	24
within a_A (aligned draws)	F1 expanded	0.6053	0.4539	+0.1514	[0.1085, 0.1959]	229 targets	24
within a_A (ordered draws)	F1 base	0.5294	0.3639	+0.1655	[0.1305, 0.2020]	229 targets	24
within a_A (ordered draws)	F1 expanded	0.6053	0.4573	+0.1480	[0.1080, 0.1905]	229 targets	24
across a_A/a_B (both directions)	CIDEr-D	1.2355	1.2191	+0.0164	[−0.0202, 0.0603]	596 targets, 107 pairs	21
across a_A/a_B (both directions)	METEOR	0.2040	0.2019	+0.0020	[−0.0016, 0.0058]	596 targets, 107 pairs	21
across a_A/a_B (both directions)	ROUGE-L	0.3307	0.3248	+0.0060	[−0.0002, 0.0134]	596 targets, 107 pairs	21

D BASELINE IMPLEMENTATION AND EVALUATION

Table 13 lists each baseline with its checkpoint, inputs, output format, decoding, and training data, and Table 14 reports overlap between that training data and our test videos.

Inputs. CLIP and SigLIP2 score a unit by the mean similarity between its representative frame and each ℓ_2-normalized interest entry of the persona card. QD-DETR and CG-DETR use each interest entry as a separate text query with their CLIP-feature checkpoints and average the resulting saliency scores. TimeChat and Qwen3-VL receive 96 uniformly sampled frames with clip-local timestamps and the full persona card. Title models receive the interval boundaries, all overlapping ASR utterances with boundary-crossing utterances marked, and the persona card, and they are asked for one factual title without forced personalization. Chapter-Llama keeps its frozen chaptering adapter and receives a single-interval naming instruction, and Vid2Seq receives frames and ASR truncated to its native 1,000-token limit.

Scoring. AP treats tied scores as a group and is undefined for empty references, which fixed masks exclude. Generated time spans outside $[0, T)$ are discarded rather than clamped, and a unit is selected when it overlaps a retained span with positive duration. Set F1 is $2|\widehat{H}_p \cap H_p|/(|\widehat{H}_p| + |H_p|)$, which equals one when a valid prediction and the reference are both empty. A prediction failure scores zero on an evaluable target, including output-limit responses whose saved text contains parseable spans. Titles are scored with the `pycocoevalcap` 1.2 implementations of CIDEr-D, METEOR, and supplementary ROUGE-L, and CIDEr-D document frequencies are fixed from 1,126 training interval documents with 8,073 reference titles.

Controls and aggregation. Within each clip or video, sorted persona identifiers are cyclically shifted to define the other persona, so other-persona outputs are the target outputs of that persona on the same content. Persona-free outputs are generated once per clip or interval and reused across target personas. Scores are averaged over intervals within a pair, personas within a clip or video, and clips within a source video, and each source video receives equal weight. Contrasts are computed from paired target-level differences before aggregation. Confidence intervals use 2,000 bootstrap resamples of source videos with seed 42, and they do not capture variation across repeated model generations. Table 15 lists generation counts and evaluation populations.

Table 13: Baselines. No model is trained on this benchmark.

Model	Checkpoint / ID	Params	Input	Persona	Output	Decoding	Training data	Notes
CLIP	openai/clip-vit-base-patch32 @3d74acf9	151.3M	representative frame per 2-s unit	interest entries	score	none (cosine, mean over entries)	general image-text pretraining	no fine-tuning
SigLIP2	google/siglip2-base-patch16-256 @3f9f96cb	375.2M	representative frame per 2-s unit	interest entries	score	none (similarity, mean over entries)	general image-text pretraining	no fine-tuning
QD-DETR	official CLIP-only ckpt (504ce581)	7.0M	CLIP ViT-B/32 features, 2-s grid, max 75 units	interest entries (one query each)	score	none (saliency, mean over queries)	QVHighlights train (CLIP-only)	queries = interest entries, saliency projected to 2-s units
CG-DETR	official CLIP-only ckpt (d51eef06)	12.0M	CLIP ViT-B/32 features, 2-s grid, max 75 units	interest entries (one query each)	score	none (saliency, mean over queries)	QVHighlights train (CLIP-only)	same query adaptation as QD-DETR
TimeChat	timechat_7b.pth (f68d52a8)	7B (foundation)	96 uniformly sampled frames	card / none	set (intervals)	sample T=0.3, top-p=0.9, max 512 new tokens	video instruction tuning (not audited)	intervals parsed and mapped to 2-s units
Qwen3-VL	Qwen/Qwen3-VL-8B-Instruct @0c351dd0	8.77B	96 ordered frames (TimeChat indices), max 16,384 px/frame	card / none	set (intervals)	sample T=0.3, top-p=0.9, max 512 new tokens	general pretraining	sensitivity run: max 2,048 new tokens
Llama-3.1-8B-Instruct	meta-llama/Llama-3.1-8B-Instruct @0e9e39f2	8B	full interval ASR	card / none	title	greedy, max 256 new tokens	general pretraining	no ASR truncation
Mistral-Nemo	mistralai/Mistral-Nemo-Instruct-2407 @04d8a905	12.25B	full interval ASR	card / none	title	greedy, max 256 new tokens	general pretraining	no ASR truncation
Qwen3-8B	Qwen/Qwen3-8B @b968826d	8.19B	full interval ASR	card / none	title	greedy, max 256 new tokens, thinking off	general pretraining	no ASR truncation
Chapter-Llama (adapted)	lucas-ventura/chapter-llama ASR LoRA @a3837e90 on Llama-3.1-8B	8.03B	full interval ASR	card / none	title	greedy, max 256 new tokens	VidChapters ASR subset (s10k-2 train, 10k videos)	one-interval naming instruction, frozen chaptering LoRA
Vid2Seq (persona-free)	GT-boundary ckpt (7bb17436)	289.1M	CLIP ViT-L/14 1 fps (100 features) + interval ASR	none	title	beam 4, max 256 new tokens	HowTo100M + VidChapters-7M	ASR cap 1,000 tokens

Table 14: Overlap between baseline training data and our test videos. Not established means that the overlap cannot be verified from released artifacts, not that the videos are unseen.

Model	Audited list	Test clips / videos in list	Sources	Other exposure
QD-DETR	QVHighlights train (supervised)	17/28 clips	17/27 sources	CLIP pretraining: not established
CG-DETR	QVHighlights train (supervised)	17/28 clips	17/27 sources	CLIP pretraining: not established
Chapter-Llama (adapted)	VidChapters ASR s10k-2 train ID list (10,000 IDs)	0/30 videos	30 sources audited	Llama-3.1 base pretraining: not established
Vid2Seq (persona-free)	VidChapters sml300 validation list	30/30 videos listed as validation	full training membership: not established	HowTo100M/CLIP/T5 membership: not established
CLIP	not audited	not established	not established	pretraining data: not established
SigLIP2	not audited	not established	not established	pretraining data: not established
TimeChat	not audited	not established	not established	pretraining data: not established
Qwen3-VL	not audited	not established	not established	pretraining data: not established
Llama-3.1-8B-Instruct	not audited	not established	not established	pretraining data: not established
Mistral-Nemo	not audited	not established	not established	pretraining data: not established
Qwen3-8B	not audited	not established	not established	pretraining data: not established

Table 15: Generation counts and evaluation populations. Prediction failures stay in the populations with score zero, whereas missing references remove a target.

Stage	Population	Units	Clips / videos	Sources	Applies to
Generation	HL, persona-conditioned	252 pairs	28 clips	27	all HL models
Generation	HL, persona-free	28 clips	28 clips	27	TimeChat, Qwen3-VL
Generation	Title, persona-conditioned	787 targets / 120 pairs	30 videos	30	title LLMs, Chapter-Llama
Generation	Title, persona-free	225 intervals	30 videos	30	title LLMs, Chapter-Llama, Vid2Seq
Evaluation	HL F1 (target, none)	242 pairs	28 clips	27	reference a_A
Evaluation	HL F1 paired (target, other, none)	239 pairs	25 clips	24	reference a_A
Evaluation	HL AP base	211 pairs	26 clips	25	reference a_A
Evaluation	HL AP expanded	223 pairs	26 clips	25	reference a_A
Evaluation	HL AP base, paired	209 pairs	24 clips	23	reference a_A
Evaluation	HL AP expanded, paired	221 pairs	24 clips	23	reference a_A
Evaluation	Title, full	294 targets / 53 pairs	16 videos	16	references a_A and a_B
Evaluation	Title, paired	275 targets / 51 pairs	14 videos	14	references a_A and a_B

E SUPPLEMENTARY EXPERIMENTAL RESULTS

Tables 16 and 17 report every condition and metric behind Tables 3 and 4, including CIDEr-D for titles. Table 18 breaks down prediction failures by model and condition, and Table 19 reports Qwen3-VL with a 2,048-token output limit, which reduces failures from 155 to 121 of 280 outputs. No paired interval is computed between the two output limits, so this run is a sensitivity check rather than a replacement. Figure 8 shows target-minus-other contrasts per source video.

Table 16: Highlight results for all conditions and metrics against the references of a_A.

		All		Paired					
Model	Metric	Target	None	Target	Other	None	T − O	95% CI	N (all / paired)
CLIP	AP base	0.1508	–	0.1384	0.1212	–	+0.0172	[0.0061, 0.0293]	211 / 209
CLIP	AP expanded	0.2302	–	0.2305	0.2134	–	+0.0171	[0.0054, 0.0307]	223 / 221
SigLIP2	AP base	0.1147	–	0.1192	0.0992	–	+0.0200	[0.0080, 0.0341]	211 / 209
SigLIP2	AP expanded	0.1831	–	0.1920	0.1784	–	+0.0136	[0.0010, 0.0276]	223 / 221
QD-DETR	AP base	0.2311	–	0.2438	0.2114	–	+0.0324	[0.0032, 0.0657]	211 / 209
QD-DETR	AP expanded	0.2987	–	0.3166	0.3014	–	+0.0152	[−0.0050, 0.0378]	223 / 221
CG-DETR	AP base	0.1975	–	0.1953	0.1605	–	+0.0349	[0.0147, 0.0608]	211 / 209
CG-DETR	AP expanded	0.2676	–	0.2753	0.2439	–	+0.0314	[0.0128, 0.0530]	223 / 221
TimeChat	F1 base	0.0854	0.0901	0.0919	0.0973	0.0974	−0.0054	[−0.0164, 0.0046]	242 / 239
TimeChat	F1 expanded	0.1754	0.1667	0.1902	0.1865	0.1736	+0.0037	[−0.0071, 0.0163]	242 / 239
TimeChat	F1 base, nonempty ref.	0.1030	0.1082	0.1076	0.1141	0.1135	−0.0065	[−0.0189, 0.0056]	211 / 209
TimeChat	F1 expanded, nonempty ref.	0.2006	0.1915	0.2106	0.2065	0.1936	+0.0041	[−0.0091, 0.0184]	223 / 221
Qwen3-VL	F1 base	0.0488	0.0923	0.0549	0.0587	0.1038	−0.0039	[−0.0136, 0.0035]	242 / 239
Qwen3-VL	F1 expanded	0.0703	0.1728	0.0791	0.0937	0.1944	−0.0145	[−0.0280, −0.0018]	242 / 239
Qwen3-VL	F1 base, nonempty ref.	0.0581	0.1129	0.0632	0.0686	0.1227	−0.0054	[−0.0187, 0.0044]	211 / 209
Qwen3-VL	F1 expanded, nonempty ref.	0.0802	0.2009	0.0872	0.1058	0.2184	−0.0186	[−0.0361, −0.0025]	223 / 221

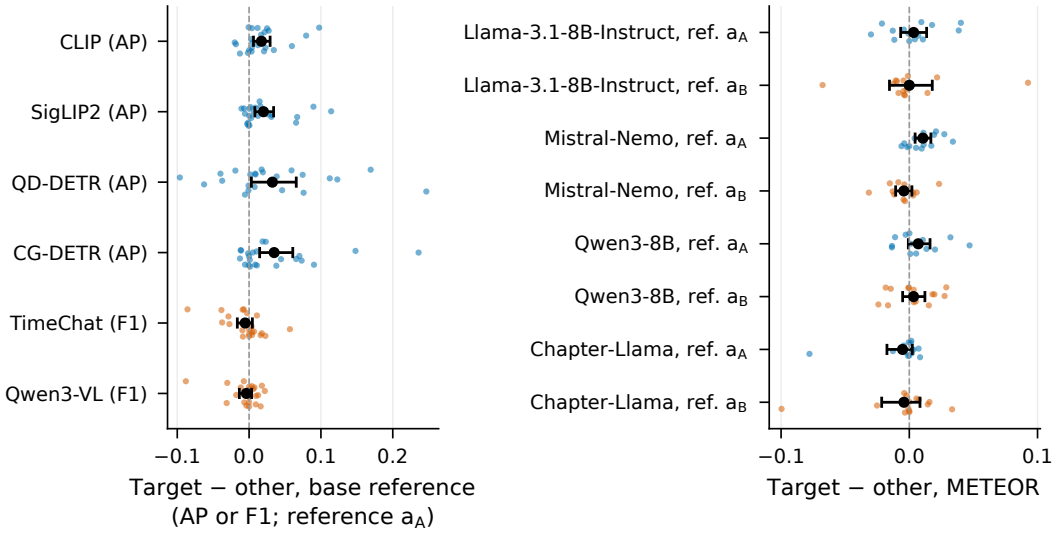


Figure 8: Target-minus-other contrasts per source video for highlight models (base references) and title models (METEOR). Black markers show the reported means and 95% intervals.

Table 17: Chapter-title results against the references of a_A and a_B. The lower block gives target-minus-other contrasts with 95% intervals. Scores against a_B are higher for all models and should be compared within a reference.

			Reference a_A			Reference a_B		
Model	Sample	Condition	CIDEr-D	METEOR	ROUGE-L	CIDEr-D	METEOR	ROUGE-L
Llama-3.1-8B-Instruct	full	target	0.8353	0.1160	0.2199	1.3621	0.1780	0.2971
Llama-3.1-8B-Instruct	full	none	0.9417	0.1305	0.2472	1.9437	0.2321	0.3652
Llama-3.1-8B-Instruct	paired	target	0.8748	0.1204	0.2301	1.4274	0.1829	0.3055
Llama-3.1-8B-Instruct	paired	other	0.8407	0.1171	0.2289	1.4300	0.1829	0.3122
Llama-3.1-8B-Instruct	paired	none	0.9416	0.1316	0.2513	1.9630	0.2351	0.3691
Llama-3.1-8B-Instruct	paired	target − other	+0.0341	+0.0034	+0.0012	−0.0026	0.0000	−0.0067
Mistral-Nemo	full	target	1.0326	0.1440	0.2581	1.5477	0.2111	0.3195
Mistral-Nemo	full	none	1.1195	0.1447	0.2779	1.8872	0.2478	0.3753
Mistral-Nemo	paired	target	1.0686	0.1463	0.2644	1.5806	0.2138	0.3241
Mistral-Nemo	paired	other	1.0049	0.1357	0.2539	1.6547	0.2181	0.3353
Mistral-Nemo	paired	none	1.1641	0.1471	0.2815	1.8968	0.2515	0.3790
Mistral-Nemo	paired	target − other	+0.0637	+0.0106	+0.0105	−0.0741	−0.0043	−0.0112
Qwen3-8B	full	target	1.1155	0.1584	0.2846	1.6797	0.2318	0.3549
Qwen3-8B	full	none	1.0819	0.1443	0.2804	1.9942	0.2476	0.3905
Qwen3-8B	paired	target	1.1157	0.1598	0.2856	1.7714	0.2391	0.3680
Qwen3-8B	paired	other	1.0744	0.1528	0.2750	1.7687	0.2358	0.3703
Qwen3-8B	paired	none	1.0943	0.1461	0.2821	2.1374	0.2576	0.4061
Qwen3-8B	paired	target − other	+0.0413	+0.0070	+0.0106	+0.0027	+0.0033	−0.0023
Chapter-Llama (adapted)	full	target	0.5671	0.0810	0.1737	1.1817	0.1375	0.2583
Chapter-Llama (adapted)	full	none	0.3787	0.0628	0.1285	1.0292	0.1155	0.2114
Chapter-Llama (adapted)	paired	target	0.5939	0.0830	0.1810	1.2615	0.1450	0.2697
Chapter-Llama (adapted)	paired	other	0.6151	0.0883	0.1889	1.3441	0.1492	0.2832
Chapter-Llama (adapted)	paired	none	0.3995	0.0653	0.1366	1.1082	0.1226	0.2261
Chapter-Llama (adapted)	paired	target − other	−0.0212	−0.0053	−0.0079	−0.0826	−0.0042	−0.0135
Vid2Seq (persona-free)	full	none	0.3997	0.0602	0.1365	1.2087	0.1211	0.2457
Vid2Seq (persona-free)	paired	none	0.3750	0.0594	0.1342	1.2403	0.1242	0.2508

Model	Reference	CIDEr-D	METEOR	ROUGE-L
Llama-3.1-8B-Instruct	a_A	+0.0341 [−0.0695, 0.1521]	+0.0034 [−0.0067, 0.0135]	+0.0012 [−0.0185, 0.0218]
Llama-3.1-8B-Instruct	a_B	−0.0026 [−0.1719, 0.1648]	0.0000 [−0.0155, 0.0179]	−0.0067 [−0.0329, 0.0181]
Mistral-Nemo	a_A	+0.0637 [0.0189, 0.1116]	+0.0106 [0.0045, 0.0168]	+0.0105 [−0.0002, 0.0205]
Mistral-Nemo	a_B	−0.0741 [−0.1637, 0.0022]	−0.0043 [−0.0107, 0.0020]	−0.0112 [−0.0272, 0.0020]
Qwen3-8B	a_A	+0.0413 [−0.0555, 0.1493]	+0.0070 [−0.0008, 0.0162]	+0.0106 [−0.0036, 0.0268]
Qwen3-8B	a_B	+0.0027 [−0.1357, 0.1214]	+0.0033 [−0.0052, 0.0122]	−0.0023 [−0.0186, 0.0123]
Chapter-Llama (adapted)	a_A	−0.0212 [−0.1057, 0.0486]	−0.0053 [−0.0174, 0.0023]	−0.0079 [−0.0353, 0.0126]
Chapter-Llama (adapted)	a_B	−0.0826 [−0.2986, 0.0673]	−0.0042 [−0.0215, 0.0083]	−0.0135 [−0.0487, 0.0095]

Table 18: Prediction outcomes by model and condition over the full test generation roster.

Model	Persona	Planned	Valid	Output limit	Format	Out of range	Parse	Persistence	Failures
CLIP	target	252	252	0	0	0	0	0	0 (0.0%)
SigLIP2	target	252	252	0	0	0	0	0	0 (0.0%)
QD-DETR	target	252	252	0	0	0	0	0	0 (0.0%)
CG-DETR	target	252	252	0	0	0	0	0	0 (0.0%)
TimeChat	target	252	250	0	0	0	2	0	2 (0.8%)
TimeChat	none	28	27	0	0	0	1	0	1 (3.6%)
Qwen3-VL	target	252	105	124	0	15	7	1	147 (58.3%)
Qwen3-VL	none	28	20	7	0	1	0	0	8 (28.6%)
Qwen3-VL (2,048 tokens)	target	252	137	70	0	37	8	0	115 (45.6%)
Qwen3-VL (2,048 tokens)	none	28	22	5	0	1	0	0	6 (21.4%)
Llama-3.1-8B-Instruct	target	787	787	0	0	0	0	0	0 (0.0%)
Llama-3.1-8B-Instruct	none	225	225	0	0	0	0	0	0 (0.0%)
Mistral-Nemo	target	787	787	0	0	0	0	0	0 (0.0%)
Mistral-Nemo	none	225	225	0	0	0	0	0	0 (0.0%)
Qwen3-8B	target	787	787	0	0	0	0	0	0 (0.0%)
Qwen3-8B	none	225	225	0	0	0	0	0	0 (0.0%)
Chapter-Llama (adapted)	target	787	599	76	112	0	0	0	188 (23.9%)
Chapter-Llama (adapted)	none	225	126	36	63	0	0	0	99 (44.0%)
Vid2Seq (persona-free)	none	225	220	0	5	0	0	0	5 (2.2%)

Table 19: Qwen3-VL with 512 and 2,048 new tokens. Only the output limit changes between the two runs.

		max 512 new tokens (main)		max 2,048 new tokens			
Metric	Condition	512	95% CI	2,048	95% CI	Paired diff. CI	N
F1 base	target	0.0488	[0.0238, 0.0788]	0.0596	[0.0320, 0.0904]	not computed	242
F1 base	none	0.0923	[0.0597, 0.1262]	0.0938	[0.0616, 0.1275]	not computed	242
F1 base	target − other (paired)	−0.0039	[−0.0136, 0.0035]	−0.0036	[−0.0150, 0.0060]	not computed	239
F1 expanded	target	0.0703	[0.0411, 0.1020]	0.0938	[0.0602, 0.1282]	not computed	242
F1 expanded	none	0.1728	[0.1185, 0.2245]	0.1772	[0.1247, 0.2275]	not computed	242
F1 expanded	target − other (paired)	−0.0145	[−0.0280, −0.0018]	−0.0145	[−0.0304, 0.0017]	not computed	239
F1 base, nonempty ref.	target	0.0581	[0.0297, 0.0895]	0.0707	[0.0406, 0.1048]	not computed	211
F1 base, nonempty ref.	none	0.1129	[0.0757, 0.1506]	0.1163	[0.0804, 0.1524]	not computed	211
F1 base, nonempty ref.	target − other (paired)	−0.0054	[−0.0187, 0.0044]	−0.0063	[−0.0203, 0.0058]	not computed	209
F1 expanded, nonempty ref.	target	0.0802	[0.0490, 0.1137]	0.1066	[0.0709, 0.1444]	not computed	223
F1 expanded, nonempty ref.	none	0.2009	[0.1403, 0.2570]	0.2066	[0.1485, 0.2619]	not computed	223
F1 expanded, nonempty ref.	target − other (paired)	−0.0186	[−0.0361, −0.0025]	−0.0193	[−0.0391, 0.0004]	not computed	221
Failures	target	147/252 (OL 124, OOR 15, parse 7, persist. 1)		115/252 (OL 70, OOR 37, parse 8)			
Failures	none	8/28 (OL 7, OOR 1, parse 0, persist. 0)		6/28 (OL 5, OOR 1, parse 0)			
Failures	all	155/280		121/280			

F Qualitative Examples and Failure Cases

We select examples by fixed hash rules rather than by scores. The highlight example is the test clip with at least three complete personas and the smallest hash of its identifier, and the title example follows the same rule over intervals with both references for at least three personas. Failure cases are the first by request-identifier hash for each failure type.

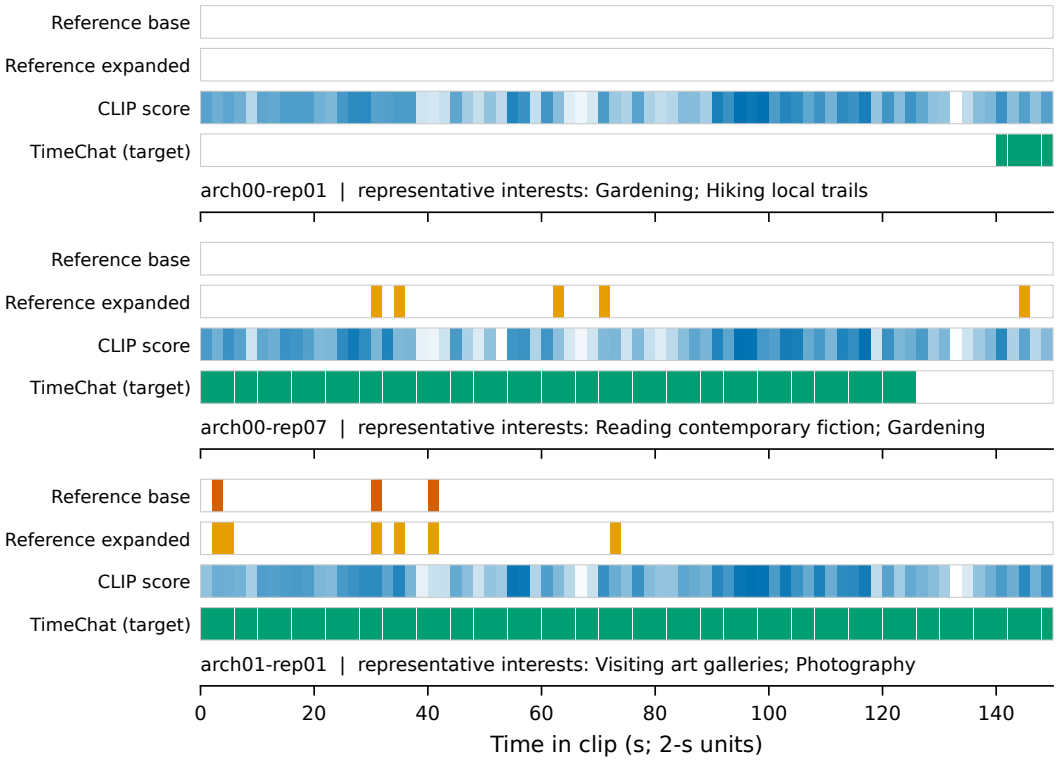


Figure 9: References of a_A, CLIP scores, and TimeChat selections for three personas on one test clip. Two personas have empty base references for this clip.

Table 20: Titles for one supplied interval under three personas. The persona-free output is shared across personas.

Persona	Reference a_A	Reference a_B	Llama-3.1 (target)	Llama-3.1 (none)
arch01-rep07	How to Receive Mail on the Road: Tips for Vans and RVs	How to Receive Mail on the Road	"Digital Illustration Techniques for the Modern Artist"	"Receiving Mail in a Van: A Guide to RV Mail Services"
arch03-rep07	Getting Mail on the Road: How RVs and Vans Handle Postal Delivery	How to Receive Mail on the Road	"Mail and Package Delivery for RVs and Vans"	"Receiving Mail in a Van: A Guide to RV Mail Services"
arch05-rep02	Getting Mail on the Road: How RVs and Vans Handle Post	How to Receive Mail on the Road	"Receiving Mail on the Road: RV and Van Mail Delivery Options"	"Receiving Mail in a Van: A Guide to RV Mail Services"

Table 21: Representative prediction failures. Outputs are truncated to 300 characters, and / marks a line break.

Model	Failure	Request	Output (first 300 characters)
Qwen3-VL (512 tokens)	output limit	`hw3xg__fQY0_60.0_210.0|arch09-rep01`	Based on the viewer's profile (84-year-old male, interests in gardening, woodworking, genealogy, church, bridge, puzzles, walking, volunteering, local history, and leisurely travel), the most interesting highlight intervals in the video are: / / 1. **0:00 - 0:06**: The opening shot of the woman speaki
Chapter-Llama (adapted)	format	`uZ9ojzYp5VU|1|__generic__`	1. Knife Sharpening
Vid2Seq (persona-free)	format	`W4CNU7JEonc|9|__generic__`	4. so you just marie kondod your life.