# Supplementary Materials
# Chapter-Llama: Efficient Chaptering in Hour-Long Videos with LLMs

Lucas Ventura[1,2] Antoine Yang[3] Cordelia Schmid[2] Gül Varol[1]

[1]LIGM, École des Ponts, IP Paris, Univ Gustave Eiffel, CNRS

[2]Inria, École normale supérieure, CNRS, PSL Research University [3]Google DeepMind

https://imagine.enpc.fr/~lucas.ventura/chapter-llama/

This appendix provides implementation details (Section A), data analysis (Section B), additional quantitative (Section C) and qualitative results (Section D). We further refer to our project page for a supplementary video visualizing the results.

## A. Implementation Details

This section provides additional implementation details for LLM finetuning (Appendix A.1), prompt structure (Appendix A.2), training data format (Appendix A.3), and the iterative prediction (Appendix A.4).

### A.1. Finetuning the LLM

As mentioned in Sec. 3, for all experiments, we finetune Llama-3.1-8B-Instruct model [4] using LoRA [6] with rank $r = 8$ and target modules Q and V projections. LoRA [6] hyperparameters are set to $\alpha = 32$ and dropout $= 0.04$. We use a batch size of 1 and a learning rate of $10^{-4}$, and train for 1 epoch using the AdamW optimizer. The training process takes 40 minutes using 4 NVIDIA H100 GPUs, and inference on 100 short videos takes 30 minutes using the same hardware.

### A.2. Prompt details

The base prompt contains the instructions as follows:

```
Given the complete transcript of a
video of duration {duration}, {task}.
Identify the approximate start time
of each chapter in the format
'hh:mm:ss - Title'.
Ensure each chapter entry is on a new
line.
Focus on significant topic changes
that would merit a new chapter in a
video, but do not provide summaries
of the chapters.
{transcript}
```

where `duration` represents the length of the video in `HH:MM:SS` format (e.g., `00:09:52`), while `task` and `transcript` are specific to the input modalities used.

For example, when utilizing both ASR and captions as input modalities, the `task` is defined as follows:

```
use the provided captions and ASR
transcript to identify distinct
chapters based on content shifts.
```

For the `transcript`, when training Chapter-Llama with both modalities, we prepend the modality names and interleave the outputs as illustrated below:

```
ASR 00:00:00: This place has blown
  our minds.
Caption 00:00:01: The image features
  two individuals, a man and a woman,
  standing outdoors in a natural
  setting with rocky terrain and
  sparse vegetation in the background.
ASR 00:00:04: Look at this.
ASR 00:00:05: In this episode, we're
  exploring Buckhorn Wash, Utah.
```

When training with only ASR (e.g., frame selector module), we simplify the input format by omitting the modality prefix, as there is only one source of information in the transcript.

We refer to Tab. A.4 for an experiment with/without these prefixes, where we observe slight gains by specifying the modalities. When using a single modality as input (e.g., ASR), there is no need to prepend the modality name to the transcript:

```
00:00:00: This place has blown
  our minds.
00:00:04: Look at this.
00:00:05: In this episode, we're
  exploring Buckhorn Wash, Utah.
```

### A.3. Training data format

For training our model, we use chapter data in the following structure. Each line contains the start timestamp of the chapter in `HH:MM:SS` format followed by the chapter title:

```
00:00:00 - We're at Buckhorn Wash,
  Utah
00:00:51 - Morrison Knudson (MK)
  Tunnels
00:01:25 - In Buckhorn Wash, Like a
  Little Zion
00:02:15 - Buckhorn Wash Pictograph
  Panel
00:03:25 - Camping in the Wash,
  Driving Through the Canyon
00:04:47 - Swinging Bridge Campground
  & San Rafael Bridge
00:06:08 - Buckhorn Draw Visitor
  Center, Well, & Spanish Trail
00:08:37 - Boondocking at Utah Lake
00:08:57 - Scenes from the Next
  Episode - Nevada: Lemoille Canyon
00:09:14 - Bloopers
```

### A.4. Iterative prediction details

As mentioned in Sec. 3 and demonstrated through experiments in Sec. 4.4 of the main paper, to handle videos with transcripts exceeding the LLM context window, we implement an iterative prediction procedure using a sliding window approach. For each video, we segment the transcript into windows of fixed token length (e.g., 20k tokens) and process them sequentially. Starting from the first window, we generate chapters for the current segment, merge them with previously generated chapters, and advance the window to the next unprocessed portion of the transcript. This process continues until the entire video is covered.

## B. Data Analysis and Statistics

Here, we provide a brief analysis of the portion from the VidChapters dataset [10] that we used in our experiments.

### B.1. Video duration distribution

Figure A.1 shows the distribution of video durations in our training set. The majority of videos (58.4%) are short videos less than 15 minutes long, while 21.9% are medium-length (15-30 minutes), 11.4% are long (30-60 minutes), and 8.3% exceed one hour. Interestingly, we observe that the average number of chapters per video increases with video duration up to about 60 minutes, where it plateaus at approximately 13 chapters. This plateau suggests a practical limit to manual chapter annotation, as annotators may be reluctant to segment videos into more than 13 chapters regardless of duration. The median video duration is 12:46 minutes.

### B.2. Video category distribution

For our final model, we use a subset of 20k training videos from VidChapters-7M. Figure A.2 compares the distribution of video categories between our training subset and the full VidChapters-7M dataset (Fig. 3 (d) [10]). As we subsample uniformly from the original training set, the two distributions closely match.

### B.3. Videos within 15k window token limit

Our models are trained with a context window of 15k tokens. In Table A.1, we analyze the breakdown of videos across categories that fall within this limit. All short and medium videos fall within this limit, while 79% of long videos also comply. Notably, for each category, the number of videos below the 15k token threshold exceeds the quantity required for model training before performance plateaus (see Fig. 4 of the main paper). This suggests that our current context window size is sufficient for effective training across all video duration categories. Note we make this analysis with the full training set of the original VidChapters dataset, as our 20k subset considers videos that 100% fall within the 15k limit.

## C. Additional Quantitative Results

We report additional results with a range of experiments, such as the impact of input and output structure (Appendix C.1, C.2, C.3), ablations with our frame selection, (Appendix C.4, C.5, C.6), the LLM training, (Ap-

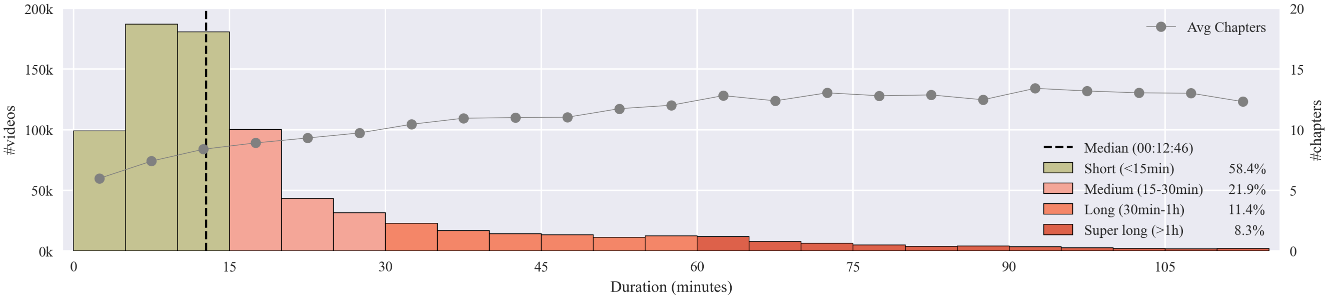


Figure A.1. **Video duration distribution:** Distribution of video durations in our training set (bars, left axis) and average number of chapters per duration bin (gray line, right axis). Most videos are less than 15 minutes long, with progressively fewer videos at longer durations. The average number of chapters increases with video duration but plateaus around 13 chapters for videos longer than one hour.

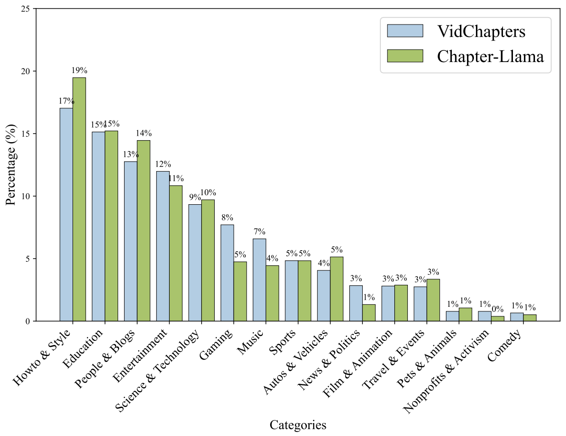


Figure A.2. **Video category distribution:** We compare the distribution of video categories between the training set of the full VidChapters-7M dataset and our 20k training subset. We observe similar distributions given our uniform sampling from the original training set.

| Category | <15k tokens | |
|---|---|---|
| Short | 466k | 100 % |
| Medium | 175k | 100 % |
| Long | 71k | 79 % |

Table A.1. **Videos in each category with fewer than 15k tokens:** We show the number of videos and proportion of short, medium, and long videos in the training set that do not exceed the 15k token limit of our training context window, from among 817k original training set videos of VidChapters. For videos without extracted captions, the caption token length are estimated by multiplying the average number of tokens per caption by the number of ground truth chapters.

pendix C.7, C.8, C.9), and further quantitative analyzes (Appendix C.10, C.11, C.12, C.13, C.14).

## C.1. Predicting timestamps without chapter titles

In our experiments, the Chapter-Llama model was trained to predict both chapter times and titles together. An alternative approach could involve training the model to predict chapter times exclusively, subsequently using another model to derive chapter titles from these times. However, as depicted in Tab. A.2, this approach underperforms compared to our current method. Therefore, we choose to continue training the Chapter-Llama model to predict both elements together, as the inclusion of chapter titles appears to enhance the accuracy of chapter time predictions.

| Ground Truth Format | F1 | tIoU | S | C |
|---|---|---|---|---|
| `HH:MM:SS` | 42.0 | 70.4 | - | - |
| `HH:MM:SS - Title` | **42.6** | **70.6** | **16.4** | **82.4** |

Table A.2. **Effect of chapter titles on timestamp prediction:** We evaluate training Chapter-Llama with only timestamps or with timestamps and chapter titles, and observe that adding chapter titles slightly improves the segmentation metrics (F1: +0.6, tIoU: +0.2).

| Modalities | | ASR | Segmentation | | Titles | |
|---|---|---|---|---|---|---|
| Speech | Capt. | timestamp | F1 | tIoU | S | C |
| ✓ | - | start end | 41.4 | 69.7 | 15.8 | 77.9 |
| | | start | 38.5 | 68.1 | 13.9 | 67.3 |
| ✓ | ✓ | start end | 39.1 | 67.6 | 6.0 | 19.9 |
| | | start | **42.6** | **70.6** | **16.4** | **82.4** |

Table A.3. **Adding end timestamps to ASR input:** Adding end timestamps to ASR transcripts improves performance when using only speech (+2.9 F1). However, when combining speech with captions, including end timestamps decreases performance significantly, especially on title metrics (e.g., 19.9 vs 82.4 CIDEr). We hypothesize this may be due to the inconsistency between modalities, where captions have single timestamps while speech segments have start and end times.

## C.2. ASR timestamp representation

As mentioned in Sec. 3, we use ASR outputs obtained with WhisperX [1], which contain start and end timestamps of each ASR segment. For our experiments, we only use the start timestamps, as opposed to using start and end timestamps of each ASR segment. In Tab. A.3, we analyze the impact of including end timestamps from ASR segments in addition to start timestamps. When using only speech inputs, including

| Has prefix? | F1 | tIoU | S | C |
|---|---|---|---|---|
| ✗ | 41.9 | 69.6 | 16.0 | 78.5 |
| ✓ | **42.6** | **70.6** | **16.4** | **82.4** |

Table A.4. **Effect of modality prefixes:** Adding prefixes to the ASR and captions modalities improves performance.

| Frame selection for captions | #frames ↓ | F1 | tIoU | S | C |
|---|---|---|---|---|---|
| Shot midpoints | 49.4 | 40.8 | 69.1 | 15.6 | 77.0 |
| Shot boundaries | 49.4 | 40.6 | 69.1 | 15.8 | 79.3 |
| Speech-based CL ±1 sec | 20.6 | **42.7** | 69.5 | **16.5** | **83.2** |
| Speech-based CL midpoints | **10.3** | 41.2 | 69.0 | 15.6 | 73.7 |
| Speech-based CL boundaries | **10.3** | 42.6 | **70.6** | 16.4 | 82.4 |

Table A.5. **Alternative frame selection strategies:** We evaluate alternative frame sampling strategies including: (1) shot boundaries and midpoints detected with `PySceneDetect` [3], (2) frames sampled ±1 second around chapter boundaries predicted by our speech-based Chapter-Llama (CL) model, (3) frames at CL predicted boundaries and midpoints between them. Results show that sampling at CL boundaries achieves competitive performance across all metrics while requiring significantly fewer frames (10.3 vs 20.6-49.4 frames per video).

end timestamps improves performance (e.g., 41.4 vs 38.5 F1). However, when training with speech and captions, using only start timestamps performs better, particularly for title generation metrics (e.g., 82.4 vs 19.9 CIDEr). We hypothesize this is because captions only have single timestamps, so having ASR segments with both start and end times creates an inconsistency between modalities that degrades performance. Therefore, in our final model we use only start timestamps for ASR segments.

## C.3. Modality prefixes

In Tab. A.4, we analyze the impact of adding modality prefixes (“`ASR:`” and “`Caption:`”) before each text segment in the interleaved input sequence. Without prefixes, the model must infer the modality type implicitly - for captions this may be easier since they often start with “The image shows”, while ASR segments have varied structure. Results show that explicitly marking modalities with prefixes improves performance across all metrics (e.g., 42.6 vs 41.9 F1), suggesting that helping the model distinguish between modalities is beneficial.

## C.4. Alternative frame selection strategies

In the main paper, given a detected chapter boundary from our speech-only model, we select frames at the boundary location itself. In Tab. A.5, we explore alternative frame sampling strategies, including: (1) shot boundaries or midpoints detected with `PySceneDetect` [3], (2) ±1 sec before and after speech-based chapter boundary predictions, (3) speech-based Chapter-Llama (CL) predicted boundary locations and midpoints between these locations. See the caption for comments.

| # videos | | Segmentation | | Titles | |
|---|---|---|---|---|---|
| F. selector | CL | F1 | tIoU | S | C |
| 1k | 1k | 42.7 | 70.8 | 15.6 | 78.1 |
| | 10k | **46.9** | **72.9** | 17.5 | 86.8 |
| 10k | 1k | 42.6 | 70.6 | 16.4 | 82.4 |
| | 10k | 46.7 | 72.2 | **18.6** | **96.4** |

Table A.6. **Effect of training data size on speech-based frame selector:** We analyze how the amount of training data used for the speech-only frame selector (first column) affects downstream performance of our Chapter-Llama (CL) model. The frame selector is trained on either 1k or 10k videos to predict frame locations where captions should be extracted, while the CL is trained on either 1k or 10k different videos for chapter generation. Comparing rows 1 vs 3 and 2 vs 4, we observe that increasing frame selector training data from 1k to 10k videos has minimal impact on segmentation metrics, but slightly improves title generation. In contrast, increasing CL training data from 1k to 10k videos (rows 1 vs 2 and 3 vs 4) improves both segmentation and title metrics.

## C.5. Training data size on the frame selection model

Throughout our experiments, we train the speech-only model using 10k videos to obtain frame locations for caption extraction (and 1k videos in most of our experiments to train our Chapter-Llama model). In Tab. A.6, we analyze how the amount of training data in the speech-only model affects downstream performance on our Chapter-Llama model using both speech and captions.

The second to last row (42.6 F1) represents our main result reported in our ablations, and the last row (46.7 F1) shows results when using 10k videos for speech-only model training and 10k videos for Chapter-Llama (CL) model training, corresponding to the final point in the *number of training videos vs performance* plot in Fig. 4 of the main paper. The first two rows show new results using only 1k videos to train the speech-only model. We observe that increasing training data for the speech-only frame selector model from 1k to 10k videos has minimal impact on segmentation metrics but improves title generation performance in both cases – from 17.5 to 18.6 SODA when using 10k videos for Chapter-Llama training, and from 15.6 to 16.4 SODA when using 1k videos for Chapter-Llama training. Increasing the training data from 1k to 10k videos for our Chapter-Llama model improves performance on both segmentation and title benchmarks, with F1 scores improving from 42.7 to 46.9 and from 42.6 to 46.7, respectively.

## C.6. Separate training data for frame selector and Chapter-Llama

In all our experiments, we use a different subset of videos to train the frame selector model and the Chapter-Llama model. In Tab. A.7, we analyze the performance of Chapter-Llama when using the same set of 1k videos for both models or when using a different set of 1k videos for the Chapter-Llama model. We see that using the same set of videos for both models decreases

| Training data | F1 | tIoU | S | C |
|---|---|---|---|---|
| $V_{F.S.} = V_{C.L.}$ | 41.4 | 70.1 | 15.1 | 77.5 |
| $V_{F.S.} \neq V_{C.L.}$ | **42.7** | **70.8** | **15.6** | **78.1** |

Table A.7. **Frame selector and Chapter-Llama training data overlap:** Given the set of videos used to train the speech-based frame selector model ($V_{F.S.}$) and and the Chapter-Llama model ($V_{C.L.}$), we compare the performance of Chapter-Llama when using different subsets of videos ($V_{F.S.} \neq V_{C.L.}$), and when using the same, already seen, videos ($V_{F.S.} = V_{C.L.}$). We see that using the same 1k set of videos for both models decreases performance.

| Llama | Speech | Captions | F1 | tIoU | S | C |
|---|---|---|---|---|---|---|
| Llama-3.2-1B | ✓ | - | 23.5 | 58.3 | 6.9 | 23.9 |
| | ✓ | ✓ | 24.6 | 58.6 | 7.4 | 28.0 |
| Llama-3.2-3B | ✓ | - | 35.2 | 66.7 | 10.5 | 52.5 |
| | ✓ | ✓ | 34.7 | 65.2 | 12.5 | 63.6 |
| Llama-3.2-11B | ✓ | - | 39.8 | 67.9 | 14.8 | 71.1 |
| | ✓ | ✓ | n/a | n/a | n/a | n/a |
| Llama-3.1-8B | ✓ | - | 38.5 | 68.1 | 13.9 | 67.3 |
| | ✓ | ✓ | **42.6** | **70.6** | **16.4** | **82.4** |

Table A.8. **Llama variants:** Model size has a significant impact on performance on Llama3.2 family. Llama-3.1-8B remains our choice due to its competitive performance with manageable computational complexity.

performance. We hypothesize that this performance drop occurs due to overfitting in the training pipeline: When both models are trained on the same videos, the outputs of the frame selector align very closely with the ground truth locations for those specific videos. This creates an artificial correlation between frame locations and content that the Chapter-Llama model learns to exploit during training. As a result, Chapter-Llama develops an over-reliance on the precise temporal positions of frames rather than learning to refine the location information.

## C.7. LLM variants

We conduct experiments with different variants of the Llama model family. All our previous results use Llama-3.1-8B-Instruct, and we now compare it against the more recent Llama-3.2 model in three sizes: 1B, 3B, and 11B parameters.

As shown in Tab. A.8, model size has a significant effect on chaptering quality. Using speech only, the F1 score improves substantially from 23.5 to 35.2 to 38.5 as we scale from 1B to 3B to 8B parameters, with only a minor additional gain to 39.8 when scaling to 11B parameters. This trend holds across all metrics. Llama-3.1-8B performs similar to Llama-3.2-11B, which we use in our final model due to reduced computational complexity. Note that we were unable to run Llama-3.2-11B on our final model combining speech and captions due to hardware constraints.

| #videos | rank | F1 | tIoU | S | C |
|---|---|---|---|---|---|
| 1k | 8 | 42.6 | 70.6 | 16.4 | 82.4 |
| | 16 | 39.9 | 68.5 | 15.6 | 78.4 |
| 5k | 8 | 45.6 | 72.3 | 18.3 | 90.0 |
| | 16 | 46.5 | 72.8 | 18.5 | 92.8 |
| 10k | 8 | 46.7 | 72.2 | 18.6 | 96.4 |
| | 16 | 46.6 | 72.4 | 18.6 | 92.5 |

Table A.9. **LoRA rank:** Comparing LoRA ranks r=8 and r=16, we find that with 1k training videos, the lower rank performs better. With 5k videos, r=16 slightly outperforms r=8. At 10k videos, both ranks achieve similar results, suggesting that with sufficient training data, model capacity becomes less important.

## C.8. LoRA rank

In Tab. A.9, we conduct experiments comparing LoRA ranks $r = 8$ and $r = 16$ across different training data sizes. With 1k training videos, the lower rank $r = 8$ performs notably better (42.6 vs 39.9 F1 score). As we increase to 5k videos, $r = 16$ shows a slight advantage (46.5 vs 45.6 F1), while at 10k videos both ranks achieve comparable performance (46.7 vs 46.6 F1). This suggests that with limited training data, a lower rank helps prevent overfitting, while with more data the model capacity becomes less critical. Based on these findings and considering efficiency, we use $r = 8$ as our default LoRA rank throughout all experiments in the paper.

## C.9. Training on videos of various durations

In most of our experiments, we have trained our model on 1k videos balanced across duration categories, i.e., 333 short videos (<15 min), 333 medium-length videos (15-30 min), and 334 long videos (30-60 min). In Tab. A.10, we show the benefit of such training on videos of various durations. For this experiment, we train new models only on 1k short videos, on 1k medium videos, and on 1k long videos. For evaluation, we use the same 300 validation videos as before, with 100 videos sampled from each duration category. As expected, training on short videos performs best on short videos (49.7 F1), while training on long videos performs best on long videos (40.4 F1). Training with a balanced mix of all three durations achieves the best overall performance across all categories (42.6 F1).

## C.10. Oracle experiments with partial ground truth input

To evaluate the Chapter-Llama model's capability in predicting chapters when provided with ground-truth chapter boundaries or titles, we conduct experiments with two scenarios: (i) incorporating ground truth timestamps into the input, and (ii) including ground truth chapter titles. In the first scenario, the task represents an upper bound limit of title metrics for our model, as it predicts chapters based on known timestamps. In the second scenario, the model predicts chapters using

| Training | Short (val) | | | | Medium (val) | | | | Long (val) | | | | All (val) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| videos | F1 | tIoU | S | C | F1 | tIoU | S | C | F1 | tIoU | S | C | F1 | tIoU | S | C |
| Short | **49.7** | **75.0** | **21.4** | **112.9** | 38.3 | 67.6 | 13.2 | 61.4 | 37.9 | 66.7 | 12.8 | 63.3 | 42.0 | 69.8 | 15.8 | 79.2 |
| Medium | 47.5 | 74.6 | 21.3 | 109.8 | 37.9 | 67.5 | 13.2 | 55.6 | 38.3 | 67.0 | 13.3 | 63.5 | 41.2 | 69.7 | 15.9 | 76.3 |
| Long | 46.6 | 74.0 | 19.5 | 104.9 | **39.3** | **68.1** | **13.4** | **62.0** | 38.1 | 66.9 | 14.3 | 75.1 | 41.3 | 69.7 | 15.8 | 80.8 |
| All | 48.4 | 74.4 | 21.2 | 110.8 | 38.9 | 68.0 | 13.1 | 57.3 | **40.4** | **69.3** | **14.9** | **79.1** | **42.6** | **70.6** | **16.4** | **82.4** |

Table A.10. **Including long videos at training improves results:** Training with 1k videos balanced across short, medium, and long durations (last row, 'All') improves performance compared to training with just 1k short videos (first row). The improvement is most pronounced for long videos (+2.5 F1). When averaging across short/medium/long validation splits, training with all videos improves all metrics: F1 (+0.6), tIoU (+0.8), S (+0.6), and C (+3.2).

| Boundaries | Titles | F1 | tIoU | S | C |
|---|---|---|---|---|---|
| ✗ | ✗ | 42.6 | 70.6 | 16.4 | 82.4 |
| ✓ | ✗ | 99.1 | 99.7 | 23.8 | 121.4 |
| ✗ | ✓ | 64.0 | 80.1 | 71.5 | 506.3 |

Table A.11. **Oracle experiment with partial ground truth input:** We evaluate the capability of Chapter-Llama in predicting chapters when provided with ground truth chapter boundaries or titles. The first scenario represents an oracle experiment for title metrics, as it predicts chapters based on known timestamps (second row). The second scenario serves as a form of video chapter grounding, i.e., given known titles to segment the boundaries (last row). The model was trained with 1k videos and evaluated with 300 videos.

| Method | F1 | tIoU | S | C |
|---|---|---|---|---|
| Vid2Seq [11] | 12.6 | 45.5 | 5.5 | 18.0 |
| Chapter-Llama (ours) | **15.5** | **49.6** | **5.0** | **26.3** |

Table A.12. **Performance on validation videos without ASR:** We evaluate the performance of our best performing model in videos without ASR predictions (190 videos in validation). We observe that the Chapter-Llama outperforms Vid2Seq in all metrics, but the performance of both models is worse than when ASR is available.

known titles, serving as a form of video chapter grounding. As demonstrated in Tab. A.11, these experiments establish the upper bounds of our model's performance.

## C.11. Performance on videos that have no speech

As mentioned in Sec. 4, most of the videos (> 97%) in the dataset have speech content. For the videos that have no ASR detections, we use every 10s sampling. We now investigate the performance of our approach when there is no ASR available. In Tab. A.12, we select all videos in the validation set without ASR, totaling 190 videos, and compare the performance to Vid2Seq [11]. We observe that the performance of both models is worse than when ASR is available, suggesting that both models mainly benefit from speech input. However, our approach still outperforms Vid2Seq in this challenging setting. By visually inspecting some of these videos, we noticed failure cases with music videos, with very similar backgrounds across frames, which makes it difficult for the model to detect chapter boundaries without any audio information. This is left to future work, as stated in the conclusions of the main paper. We also notice success cases often depict frames with text, which are captured by the captioner (see first and last examples in Fig. A.7).

## C.12. Full set of metrics

In Sec. 4.1 of the main paper, we adopted the evaluation metrics (F1, tIoU, SODA, and CIDEr), which we consider more suitable for assessing video chapter generation. For completeness and direct comparison with VidChapters [10], we also report results using their full set of metrics in Tabs. A.13 and A.14. The segmentation metrics include precision and recall at 3-second and 5-second thresholds, as well as at 0.5 and 0.7 IoU thresholds. The full metrics (referred to as 'global metrics' by [10]) comprise SODA (S) [5], BLEU (B1-B4) [8], CIDEr (C) [9], METEOR (M) [2], and ROUGE-L (RL) [7]. Our model consistently outperforms Vid2Seq [11] across all metrics.

## C.13. Repetition analysis

We have noticed that Vid2Seq tends to repeat chapter titles (see Fig. 3 of the main paper). To quantify this, we calculate the ratio of unique chapter titles to the total number of chapter titles predicted for each video and then average this ratio across all videos in the test set. For the ground truth, this average ratio is 99.6%, i.e., almost all chapter titles are unique. For our finetuned model, this average ratio is 96.3%. In contrast, Vid2Seq has a much lower average ratio of 63.5%, indicating that it indeed repeats chapter titles frequently.

## C.14. Accuracy of number of chapter predictions

While our main evaluation focused on the quality of chapter segment predictions, it is also important to assess the accuracy in predicting the number of chapters. Our primary metrics (F1, tIoU, SODA, and CIDEr) do not directly indicate whether the predicted chapter count is correct or if the method tends to over- or under-segment. To evaluate this, we analyze the distribution of differences between predicted and ground truth chapter counts for Chapter-Llama, Zero-shot, and Vid2Seq models, as illustrated in Fig. A.3.

The results reveal that Chapter-Llama exhibits the most concentrated distribution centered around zero, indicating

| Method | P@5s | R@5s | P@3s | R@3s | P@0.5 | R@0.5 | P@0.7 | R@0.7 |
|---|---|---|---|---|---|---|---|---|
| Vid2Seq [11] | 30.6 | 36.4 | 24.4 | 28.7 | 46.3 | 51.1 | 28.7 | 30.6 |
| Chapter-Llama | **52.0** | **51.7** | **45.1** | **44.7** | **66.3** | **63.4** | **49.9** | **47.8** |

Table A.13. **Video chapter generation (segmentation metrics) on VidChapters [10] test set:** Comparison of segmentation metrics between Vid2Seq and our best model from Tab. 1. Metrics include precision and recall at 3-second and 5-second thresholds, as well as at 0.5 and 0.7 IoU thresholds. Our method consistently outperforms Vid2Seq across all metrics.

| Method | S | B1 | B2 | B3 | B4 | C | M | RL |
|---|---|---|---|---|---|---|---|---|
| Vid2Seq [11] | 11.6 | 11.1 | 7.7 | 4.5 | 3.1 | 55.8 | 9.6 | 12.8 |
| Chapter-Llama | **19.3** | **19.5** | **14.3** | **8.7** | **5.6** | **100.9** | **15.4** | **22.2** |

Table A.14. **Full metrics used by VidChapters [10]:** We report the full metrics (referred to as 'global metrics' in [10]) on the test set of VidChapters. We compare Vid2Seq and our best model from Tab. 1. Metrics include SODA [5] (S), BLEU [8] (B1-B4), CIDEr [9] (C), METEOR [2] (M), and ROUGE-L [7] (RL). Our method consistently outperforms Vid2Seq across all metrics.

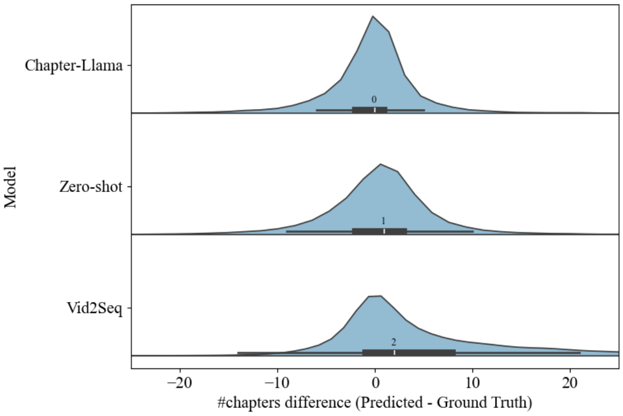


Figure A.3. **Accuracy of number of chapter predictions:** The violin plot shows the distribution of differences between the predicted and ground truth number of chapters for three video chaptering models: Chapter-Llama, Zero-shot, and Vid2Seq. The Chapter-Llama model exhibits the most concentrated distribution centered around 0, indicating accurate number of chapter prediction. The Zero-shot model tends to slightly overpredict the number of chapters, while the Vid2Seq model often significantly overpredicts the number of chapters. The median differences are 0, 1, and 2 for Chapter-Llama, Zero-shot, and Vid2Seq, respectively, with mean number of chapter differences of -0.2, 0.5, and 4.5 (not shown).

superior accuracy in predicting chapter counts. In contrast, both Zero-shot and Vid2Seq models over-segments the video with a high number of chapters. The tight interquartile range and symmetrical density shape of Chapter-Llama suggest a more reliable chapter count prediction. However, it is important to note that accurately predicting the number of chapters does not necessarily guarantee correct chapter segmentation.

# D. Additional Qualitative Analyses

We present several qualitative analyses: (i) evaluation metric calculation examples (Appendix D.1), (ii) caption visualizations (Appendix D.2), and (iii) predictions from our model (Appendix D.3).

## D.1. Evaluation metrics

In Sec. 4.1, we introduced our primary evaluation metrics for video chaptering: **tIoU** and **F1** scores. Here, we illustrate how these metrics are calculated using concrete examples, as shown in Fig. A.4.

For tIoU (temporal Intersection over Union), we first match predicted and ground truth segments by greedily selecting pairs with the highest IoU scores. In the top example of Fig. A.4, we have 5 ground truth chapters and 4 predicted chapters. The matching process starts with chapters having the most overlap, and each chapter can only be used once. The tIoU score (84.7) is then calculated as the mean IoU across all matched pairs (97.6, 53.6, 89.3, 98.3). Similarly, for the bottom example, the tIoU score of 49.4 is the mean of 60.7, 47.14, and 40.3.

For the F1 score, we compute precision and recall at different IoU thresholds (from 0.5 to 0.95 with a step of 0.05). In the top example, at a threshold of 0.5, all predicted chapters have a ground truth match with an overlap higher than 50%, resulting in a precision of 100%. However, one ground truth chapter out of 5 is left without a prediction, leading to a recall of 80%. The F1 score is then computed as the harmonic mean of precision and recall. This process is repeated for all thresholds, and the final F1 metric is the average across these thresholds.

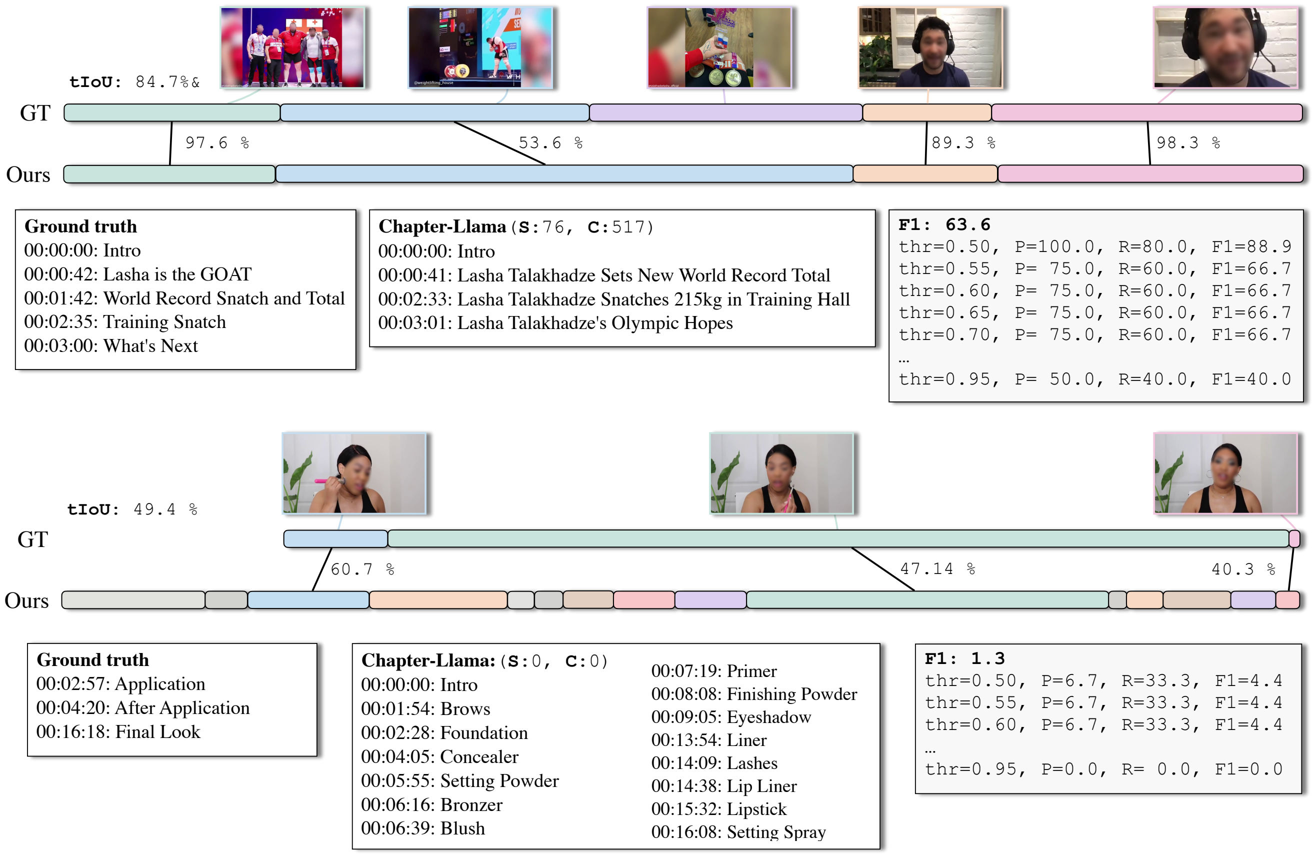
## D.2. Visualizing captions

In Fig. A.5, we provide an example, where we also visualize some of the intermediate captions that are fed to our chapter generation LLM. We then show the chapter predictions from the speech-based frame selection model, the corresponding captions selected based on this model, and the refined predictions with Chapter-Llama.

## D.3. Chapter-Llama prediction examples

Similar to Fig. 3 of the main paper, in Fig. A.6, we present two additional examples comparing our method against Vid2Seq and our zero-shot baseline.

In Fig. A.7, we show three examples of our Chapter-Llama predictions compared to the ground truth (GT) for videos without speech (3% of the data). We observe that many of the



**Ground truth**
00:00:00: Intro
00:00:42: Lasha is the GOAT
00:01:42: World Record Snatch and Total
00:02:35: Training Snatch
00:03:00: What's Next

**Chapter-Llama** (S:76, C:517)
00:00:00: Intro
00:00:41: Lasha Talakhadze Sets New World Record Total
00:02:33: Lasha Talakhadze Snatches 215kg in Training Hall
00:03:01: Lasha Talakhadze's Olympic Hopes

```
F1: 63.6
thr=0.50, P=100.0, R=80.0, F1=88.9
thr=0.55, P= 75.0, R=60.0, F1=66.7
thr=0.60, P= 75.0, R=60.0, F1=66.7
thr=0.65, P= 75.0, R=60.0, F1=66.7
thr=0.70, P= 75.0, R=60.0, F1=66.7
...
thr=0.95, P= 50.0, R=40.0, F1=40.0
```

**Ground truth**
00:02:57: Application
00:04:20: After Application
00:16:18: Final Look

**Chapter-Llama:** (S:0, C:0)
00:00:00: Intro
00:01:54: Brows
00:02:28: Foundation
00:04:05: Concealer
00:05:55: Setting Powder
00:06:16: Bronzer
00:06:39: Blush
00:07:19: Primer
00:08:08: Finishing Powder
00:09:05: Eyeshadow
00:13:54: Liner
00:14:09: Lashes
00:14:38: Lip Liner
00:15:32: Lipstick
00:16:08: Setting Spray

```
F1: 1.3
thr=0.50, P=6.7, R=33.3, F1=4.4
thr=0.55, P=6.7, R=33.3, F1=4.4
thr=0.60, P=6.7, R=33.3, F1=4.4
...
thr=0.95, P=0.0, R= 0.0, F1=0.0
```

Figure A.4. **Segmentation metrics visualization:** We illustrate with examples how tIoU and F1 scores are calculated for video chaptering. The top example shows a high-quality prediction with good overlap, while the bottom example demonstrates a lower-quality prediction with more misalignments. We additionally show the corresponding SODA (S) and CIDEr (C) scores.

completely ‘speechless’ videos contain OCR-readable text to help the viewer follow the video (top and bottom examples), in which cases the captioners tend to perform OCR, leading to satisfactory chaptering results. Otherwise, in case of no on-screen text and no speech (e.g., only music), the result is inferior, though still acceptable (middle example). As also evaluated in Tab. A.12, our model still achieves reasonable quantitative performance, even if speech indeed tends to be more informative for chaptering than visual modality [10].

## References

[1] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. In *Interspeech*, 2023. 3

[2] 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*, 2005. 6, 7

[3] Brandon Castellano. Pyscenedetect: Intelligent scene cut detection and video splitting tool. https://pyscenedetect.readthedocs.io/en/latest/, 2018. 4

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv:2407.21783*, 2024. 1

[5] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. SODA: Story oriented dense video captioning evaluation framework. In *ECCV*, 2020. 6, 7

[6] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 1

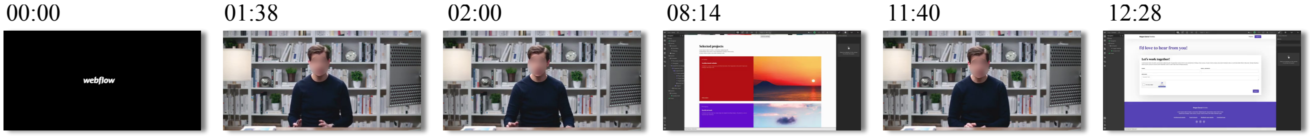
[7] Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS)*, 2004. 6, 7

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6, 7

[9] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 6, 7

[10] Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. VidChapters-7M: Video chapters at scale. In *NeurIPS Track on Datasets and Benchmarks*, 2023. 2, 6, 7, 8

[11] 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 *CVPR*, 2023. 6, 7



**Ground truth**
00:00: Day 12, begin
01:38: Homepage
09:14: Project pages
11:40: Contact page
12:28: Recap

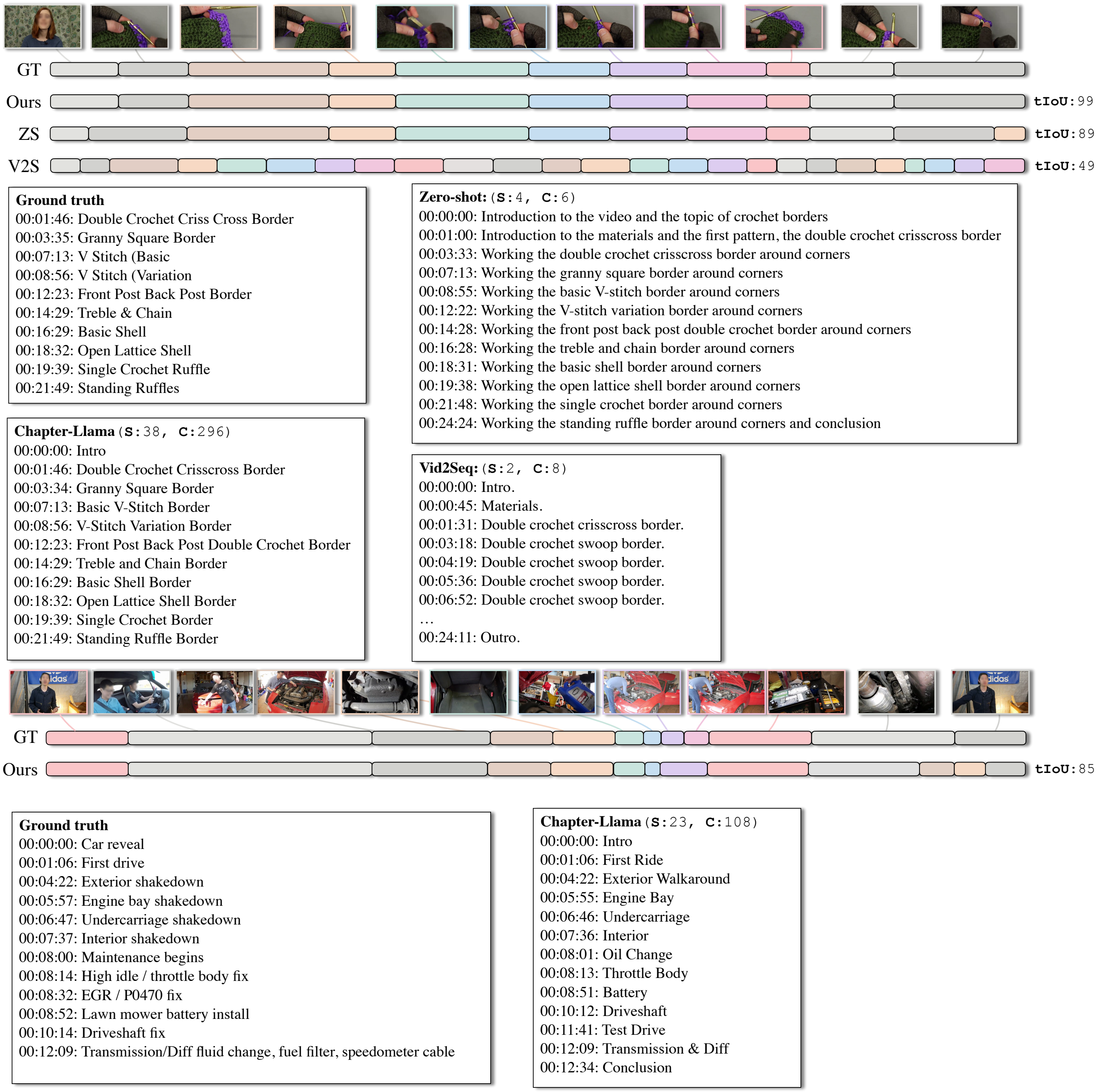
**Frame selector** (`S: 49, C: 187`)
00:00: Introduction
01:38: Backup
02:00: Home Page
08:14: Project Pages
11:40: Contact Page
12:28: Recap

**Chapter-Llama** (`S: 54, C: 225`)
00:00: Introduction
01:38: Light Mode
08:14: Project Pages
11:40: Contact Page
12:28: Recap

**Captions**

- 00:00: The image features the word "webflow" written in white text against a black background.
- 01:38: The image depicts a man standing in front of a desk with a large bookshelf behind him. The bookshelf is filled with numerous books, suggesting that the setting might be a library or a study area. The man appears to be engaged in a conversation or giving a presentation, as he is gesturing with his hands. There are also some potted plants and a clock visible in the background, adding to the ambiance of the room.
- 02:00: The image depicts a man sitting at a desk in a room filled with bookshelves. He appears to be in a library or a study area. The man is wearing a dark-colored shirt and is engaged in a conversation, possibly giving a presentation or discussing something with someone. The room is well-lit, and there are various items around the man, including a clock, a potted plant, and a few books on the shelves.
- 08:14: The image shows a screenshot of a webpage or application interface, likely a project management or portfolio platform, where a user can view and manage their selected projects.
- 11:40: The image depicts a man standing in front of a desk in a room filled with bookshelves. The man appears to be in a library or a study area, as there are numerous bookshelves surrounding him. He is wearing a dark-colored sweater and is engaged in a conversation, possibly giving a presentation or discussing a topic. The room also features a clock on the wall, adding to the academic or intellectual ambiance of the setting.
- 12:28: The image shows a screenshot of a webpage with a purple and white color scheme. The main text on the page is in a large, bold font and reads "I'd love to hear from you! Let's work together!" The page also includes a form with a purple background and a purple button that says "Let's work together!" The user interface suggests that the webpage is likely a form for contacting or working with someone, possibly a business or service provider.

Figure A.5. **Visualizing captions:** We provide an example with chapter predictions using the speech-based frame selection model, the corresponding captions sampled, and the refined predictions produced by Chapter-Llama. We additionally show the corresponding SODA (S) and CIDEr (C) scores. We see that the initially predicted chapter at timestamp 02:00 is suppressed by Chapter-Llama.



**Ground truth**
00:01:46: Double Crochet Criss Cross Border
00:03:35: Granny Square Border
00:07:13: V Stitch (Basic
00:08:56: V Stitch (Variation
00:12:23: Front Post Back Post Border
00:14:29: Treble & Chain
00:16:29: Basic Shell
00:18:32: Open Lattice Shell
00:19:39: Single Crochet Ruffle
00:21:49: Standing Ruffles

**Chapter-Llama** (S:38, C:296)
00:00:00: Intro
00:01:46: Double Crochet Crisscross Border
00:03:34: Granny Square Border
00:07:13: Basic V-Stitch Border
00:08:56: V-Stitch Variation Border
00:12:23: Front Post Back Post Double Crochet Border
00:14:29: Treble and Chain Border
00:16:29: Basic Shell Border
00:18:32: Open Lattice Shell Border
00:19:39: Single Crochet Border
00:21:49: Standing Ruffle Border

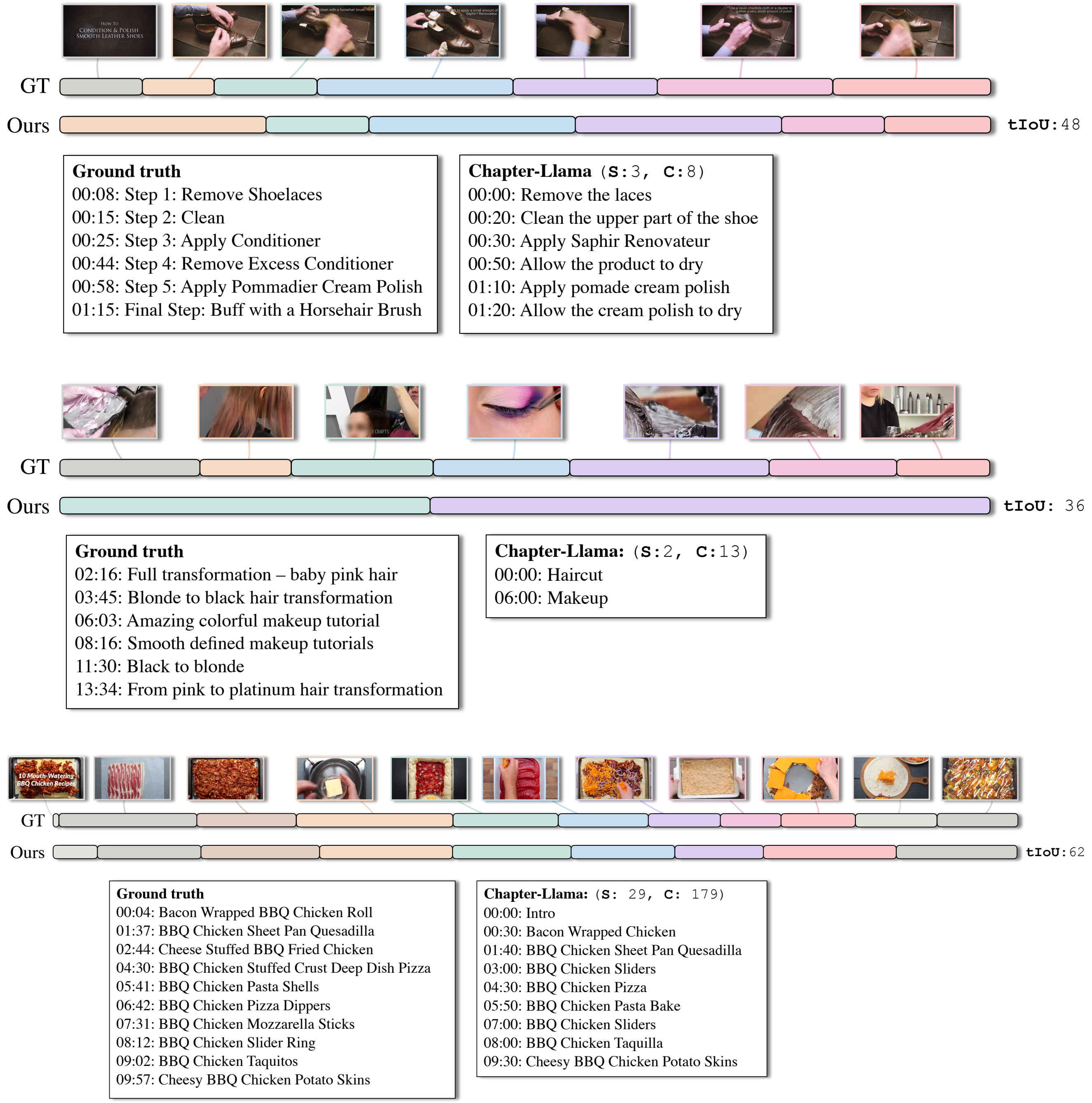
**Zero-shot:** (S:4, C:6)
00:00:00: Introduction to the video and the topic of crochet borders
00:01:00: Introduction to the materials and the first pattern, the double crochet crisscross border
00:03:33: Working the double crochet crisscross border around corners
00:07:13: Working the granny square border around corners
00:08:55: Working the basic V-stitch border around corners
00:12:22: Working the V-stitch variation border around corners
00:14:28: Working the front post back post double crochet border around corners
00:16:28: Working the treble and chain border around corners
00:18:31: Working the basic shell border around corners
00:19:38: Working the open lattice shell border around corners
00:21:48: Working the single crochet border around corners
00:24:24: Working the standing ruffle border around corners and conclusion

**Vid2Seq:** (S:2, C:8)
00:00:00: Intro.
00:00:45: Materials.
00:01:31: Double crochet crisscross border.
00:03:18: Double crochet swoop border.
00:04:19: Double crochet swoop border.
00:05:36: Double crochet swoop border.
00:06:52: Double crochet swoop border.
…
00:24:11: Outro.

**Ground truth**
00:00:00: Car reveal
00:01:06: First drive
00:04:22: Exterior shakedown
00:05:57: Engine bay shakedown
00:06:47: Undercarriage shakedown
00:07:37: Interior shakedown
00:08:00: Maintenance begins
00:08:14: High idle / throttle body fix
00:08:32: EGR / P0470 fix
00:08:52: Lawn mower battery install
00:10:14: Driveshaft fix
00:12:09: Transmission/Diff fluid change, fuel filter, speedometer cable

**Chapter-Llama** (S:23, C:108)
00:00:00: Intro
00:01:06: First Ride
00:04:22: Exterior Walkaround
00:05:55: Engine Bay
00:06:46: Undercarriage
00:07:36: Interior
00:08:01: Oil Change
00:08:13: Throttle Body
00:08:51: Battery
00:10:12: Driveshaft
00:11:41: Test Drive
00:12:09: Transmission & Diff
00:12:34: Conclusion

Figure A.6. **Additional qualitative examples:** We show two more examples of our Chapter-Llama predictions compared to the ground truth (GT). Our method generates accurate temporal boundaries and relevant chapter titles that align well with the video content. For each example, we display the corresponding SODA (S) and CIDEr (C) scores.



**Ground truth**
00:08: Step 1: Remove Shoelaces
00:15: Step 2: Clean
00:25: Step 3: Apply Conditioner
00:44: Step 4: Remove Excess Conditioner
00:58: Step 5: Apply Pommadier Cream Polish
01:15: Final Step: Buff with a Horsehair Brush

**Chapter-Llama** (**S**:3, **C**:8)
00:00: Remove the laces
00:20: Clean the upper part of the shoe
00:30: Apply Saphir Renovateur
00:50: Allow the product to dry
01:10: Apply pomade cream polish
01:20: Allow the cream polish to dry

**Ground truth**
02:16: Full transformation – baby pink hair
03:45: Blonde to black hair transformation
06:03: Amazing colorful makeup tutorial
08:16: Smooth defined makeup tutorials
11:30: Black to blonde
13:34: From pink to platinum hair transformation

**Chapter-Llama:** (**S**:2, **C**:13)
00:00: Haircut
06:00: Makeup

**Ground truth**
00:04: Bacon Wrapped BBQ Chicken Roll
01:37: BBQ Chicken Sheet Pan Quesadilla
02:44: Cheese Stuffed BBQ Fried Chicken
04:30: BBQ Chicken Stuffed Crust Deep Dish Pizza
05:41: BBQ Chicken Pasta Shells
06:42: BBQ Chicken Pizza Dippers
07:31: BBQ Chicken Mozzarella Sticks
08:12: BBQ Chicken Slider Ring
09:02: BBQ Chicken Taquitos
09:57: Cheesy BBQ Chicken Potato Skins

**Chapter-Llama:** (**S**: 29, **C**: 179)
00:00: Intro
00:30: Bacon Wrapped Chicken
01:40: BBQ Chicken Sheet Pan Quesadilla
03:00: BBQ Chicken Sliders
04:30: BBQ Chicken Pizza
05:50: BBQ Chicken Pasta Bake
07:00: BBQ Chicken Sliders
08:00: BBQ Chicken Taquilla
09:30: Cheesy BBQ Chicken Potato Skins

Figure A.7. **Additional qualitative examples without ASR:** We show three examples of videos without speech, comparing our Chapter-Llama predictions to ground truth (GT). Despite lacking ASR, our method still produces reasonable chapters by leveraging visual cues and on-screen text when available (top and bottom examples). For each example, we display the corresponding SODA (S) and CIDEr (C) scores.