# HiVid-Narrator: Hierarchical Video Narrative Generation with Scene-Primed ASR-anchored Compression

Supplementary Material

## 6. Dataset Statistics

Figure 4 shows the category distribution of E-HVC-Bench, which spans 13 e-commerce categories with relatively balanced coverage. To provide a more comprehensive overview of the datasets used in our study, Table 7 further compares our proposed E-HVC dataset (including the large-scale training set E-HVC-146K and the evaluation benchmark E-HVC-Bench) with several existing public video analysis datasets.

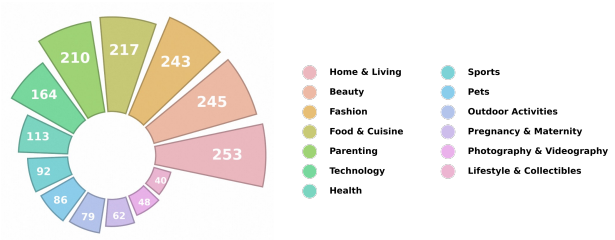


Figure 4. E-HVC-Bench: Distribution of 1,852 benchmark videos across 13 categories.

The statistics include the number of videos, the average number of temporal segments or chapters per video (referred to as 'Chapters/Video', which corresponds to the average number of clips per video as defined in the main paper), and the average text length metrics. For datasets that do not feature a dual-granularity annotation scheme like E-HVC (i.e., they lack a 'Temporal Chain-of-Thought' equivalent), the 'Think Length' columns are marked with -. The 'Answer Length' columns represent the average length of the main descriptive or summary text per video and per second. For E-HVC, these correspond to the average length of the 'Chapter Summary' component. The unit for text length is characters, and the language is indicated in parentheses where known.

As detailed in the main paper, the E-HVC dataset is unique in providing dual-granularity annotations: the Temporal Chain-of-Thought for fine-grained event-level understanding and the Chapter Summary for coherent, high-level narrative structuring. This is reflected in the statistics, where E-HVC datasets have values for both 'Think' and 'Answer' length metrics, while other datasets only have 'Answer' metrics. The statistics confirm that E-HVC-146K is a large-scale dataset, and both E-HVC-146K and E-HVC-Bench contain significantly more detailed text descriptions (both in the 'Think' and 'Answer' components) compared to the baseline public datasets, reflecting the rich, structured nature of our annotations tailored for e-commerce video narrative generation.

## 7. Compression Ratio Analysis for SPA-Compressor

This section provides a detailed analysis of the compression ratios achieved by our proposed Scene-Primed ASR-anchored Compressor (SPA-Compressor). The core idea is to compress the high-dimensional visual tokens generated by the backbone vision encoder into a more compact, hierarchical representation consisting of scene-level ($S$) and event-level ($E$) tokens.

The original input sequence for a video often contains a large number of visual tokens, which dominate the sequence and create computational overhead due to the quadratic complexity of transformer self-attention. The SPA-Compressor addresses this by leveraging ASR transcripts to segment the video into sentence-aligned segments. For each segment, it compresses the corresponding visual tokens into $S$ scene tokens and $N \times E$ event tokens, where $N$ represents the average number of image frames associated with each ASR sentence. This results in a total of $S + N \times E$ compressed visual tokens per ASR segment, significantly reducing the input length.

The compression ratio for each ASR segment is calculated as:

$$\text{Compression Ratio} = \frac{S + N \times E}{N \times D_v} = \frac{S/N + E}{D_v} \tag{13}$$

where $D_v$ is the dimensionality of the original visual tokens (e.g., $D_v = 384$ for siglip-so400m-patch 14-384).

The token reduction percentage is calculated as:

$$\text{Token Reduction (\%)} = (1 - \text{Compression Ratio}) \times 100 \tag{14}$$

Our analysis of the E-HVC dataset reveals that $N$, the average number of image frames per ASR sentence, is approximately 1.836. For the specific configuration used in our main experiments, $S = 64$ and $E = 32$, the compression ratio is calculated as:

$$\text{Compression Ratio} = \frac{64 + 1.836 \times 32}{1.836 \times 384} \tag{15}$$

$$= \frac{64 + 58.752}{704.832} \tag{16}$$

$$= \frac{122.752}{704.832} \approx 0.1741 \tag{17}$$

| Dataset | # Videos | Chapters/Video | Think Length/Video | Think Length/s | Answer Length/Video | Answer Length/s |
|---|---|---|---|---|---|---|
| YouCook2 | 2K | 7.7 | - | - | 67.7 (en) | 0.21 |
| ActivityNet Captions | 20K | 3.7 | - | - | 47.6 (en) | 0.40 |
| Charades-STA | 10K | 1.8 | - | - | 11.0 (en) | 0.36 |
| ViTT | 8K | 5.0 | - | - | 110.5 (en) | 0.44 |
| VideoStory | 20K | 6.1 | - | - | - | - |
| Video Storytelling | 105 | 13.5 | - | - | 162.6 (en) | 0.22 |
| E-SyncVidStory | 6K | 6.9 | - | - | 194.1 (zh) | 5.21 |
| E-HVC-146K | 146K | 4.1 | 693.0 | 17.0 | 290.0(zh) | 6.94 |
| E-HVC-Bench | 1.8K | 4.2 | 706.8 | 17.4 | 309.1(zh) | 7.42 |

Table 7. Comparison of video analysis datasets. For datasets without think sections, corresponding columns are marked with '-'. Text lengths are in characters; language is indicated in parentheses where applicable. The 'Chapters/Video' column represents the average number of clips per video.

This corresponds to a token reduction of $(1 - 0.1741) \times 100 \approx 82.59\%$. This aligns closely with the 82.59% token reduction reported in the main paper (Table 6), confirming the high efficiency of our SPA-Compressor in achieving significant input token reduction while maintaining essential visual details through the hierarchical scene-event modeling.