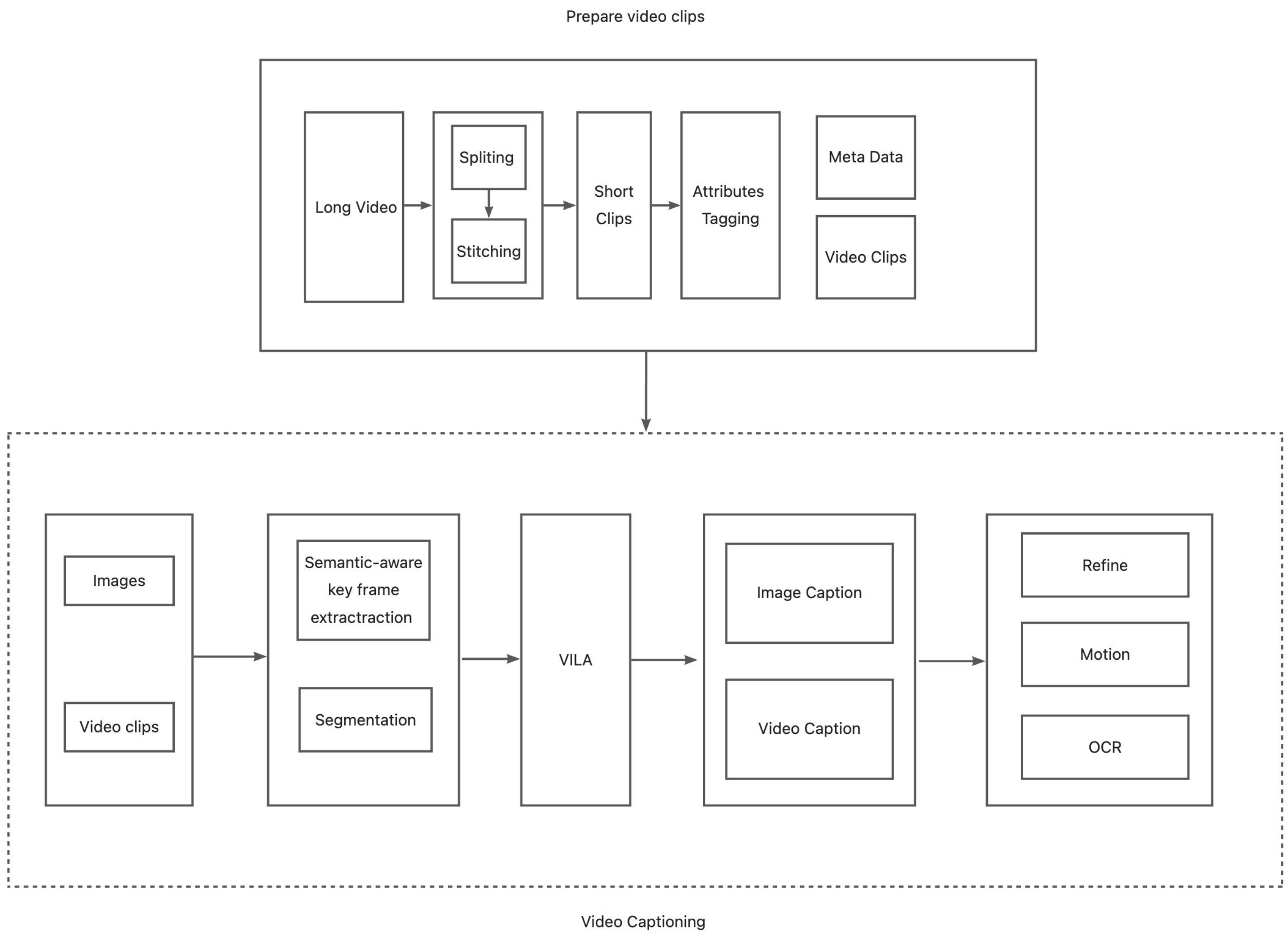
Curation Pipeline

Challenges of designing a high-quality video captioning

1. Inter-frame precise temporal change understanding. (帧间精确的时间变化理解)
2. Intra-frame detailed content description. （帧内详细内容描述）
3. Frame-number scalability for arbitrary-length videos. (任意长度视频的帧数可扩展性)

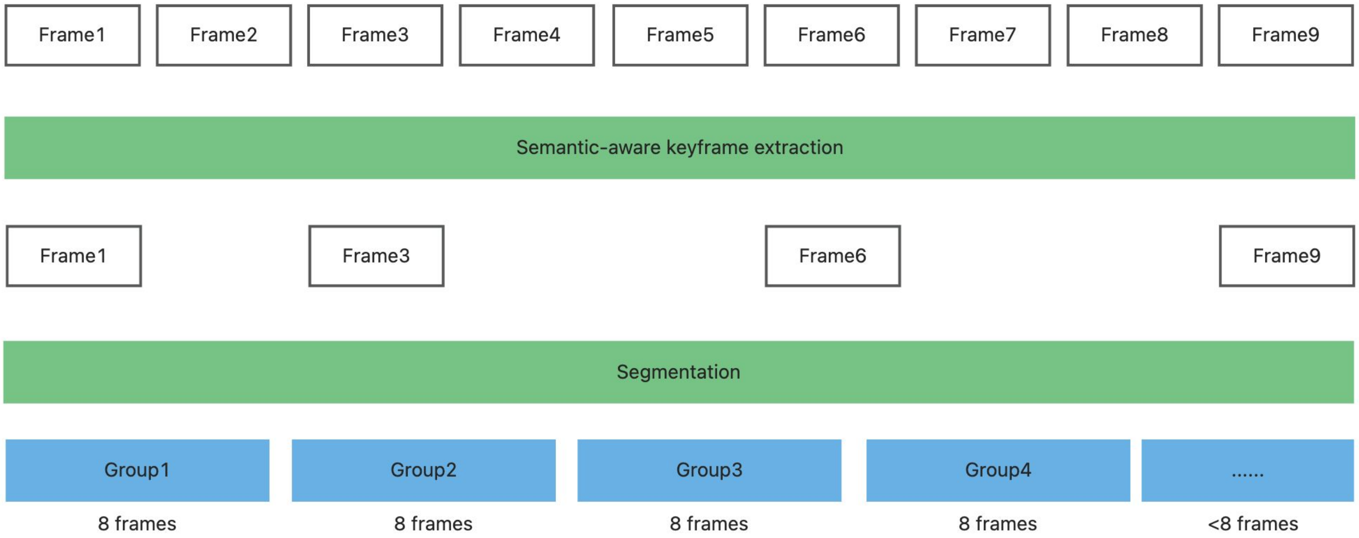
The final video dataset should contains high-quality videos spanning a wide range of categories, and the resulting captions encompass rich world knowledge, object attributes, camera movements, and crucially, detailed and precise temporal descriptions of events. (最终视频数据集应该包含广泛的类别，其字幕包括丰富的世界知识、对象属性、相机移动，以及重要的事件的详细描述和精确的时间描述)

Pipeline



Preprocess

- Semantic-aware key frame extraction: Applying semantic-aware keyframe extraction to sparse sampling to maintain significant semantic change (将语义感知关键帧提取用于视频帧的稀疏采样，使得采样帧之间保持显著的语义变化)



```python
# Input:  video,              - Video to be processed
#         time_interval,      - Minimum time interval between keyframes
#         threshold           - Inter-frame similarity threshold

# Output: key_frame_pool - Extracted keyframes
# Minimum frame interval
frame_interval = time_interval * video.fps
# All frames of the video
video_frame_list = video.frame_list

# Initialize the keyframe pool using the first frame of the video
key_frame_pool.append(video_frame_list[0])

# Semantic-aware keyframe extract
for idx in range(1, len(video_frame_list), frame_interval):
    frame_feature_1 = get_clip_cls_token(key_frame_pool[-1])
    frame_feature_2 = get_clip_cls_token(video_frame_list[idx])
    similarity = similarity(frame_feature_1, frame_feature_2)
    if similarity < threshold:
        key_frame_pool.append(video_frame_list[idx])
# Accommodate videos with small content changes
if len(key_frame_pool) == 1:
    key_frame_pool.append(video_frame_list[-1])
```

Captioning

Prompt Template

- Caption Prompt Template

```Scheme
# System
Please provide clear, sequential descriptions for the given frames. You excel in
identifying and conveying changes in actions, behaviors, envrionment, states and
attributes of objects, and camera movements between adjacent video frames.

# Skills
1. Describing object actions and behaviors
- Describe the action or behavior of objects within the frame.
2. Describing environment and background variations
- Elaborate on environment and background alterations seen between frames.
3. Describing object appearances
- Describe the appearance of objects within the frame.
4. Describing camera movements
- Perceive the camera's movements, such as panning or zooming.

# Constraints
- State facts objectively without using any rhetorical devices such as metaphors
or personification.
- Stick to a narrative format for descriptions, avoding list-like itemizations.
- Descriptions need to be concise, describing only the information that can be
determined, without analysis or speculation.
```

- Summary Prompt Template

```Scheme
# System
You will be provided with the descriptions of each of mutiple consecutive frames
in a video, each containing the content of the current multiple frames and how the
current multiple frames have changed relative to the previous frames. Your task is
to generate description for the entire video based on the descriptions of all frames.

# Skills
- Summarize sequentially, maintaining coherence between frames and the integrity
of the timeline.

# Constraints
- Don't analyze, subjective interpretations, aesthetic rhetorc, etc., just
object statements.
- Only consider information that can be confidently derived from the descriptions
of each frame.
- Do not extrapolate or imagine, remove uncertain information.
```

Postprocess

- Motion, Ocr, Duration filter
- Caption Refine
  - Special character cleaning
  - Redundant words filtering

```scheme
[
    "The video shows",
    "The video captures",
    "The video features",
    "The video depicts",
    "The video presents",
    "The video features",
    "The video is ",
    "In the video,",
    "The image shows",
    "The image captures",
    "The image features",
    "The image depicts",
    "The image presents",
    "The image features",
    "The image is ",
    "The image portrays",
    "In the image,",
]
```

HyperSora Data Recipe

- V2.1

Video clips:

Source	Samples(2–20, OCR, motion, b>=20)	Ratio
pexels	133,586	3.11%
freepik	33,853	0.79%
pixabay	37,484	0.87%
mixkit	18,455	0.43%
netflix	524,460	12.19%
mira(0.7)	311,609	7.24%
iqiyi(0.5)	119,852	2.79%
bt	733,054	17.03%

youtube	2,389,901	55.55%
Total	4,302,254	

Images

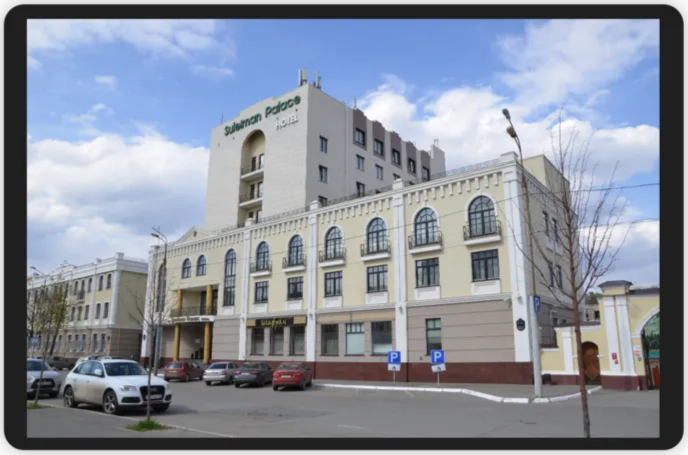
Source	Samples
Laion	1M
SAM	5M
Total	6M

Examples

7a786acfec87bfabcea098df8dcd9193.mp4

Plain Text

```
The video begins with a bird's eye view of a busy city street with people walking
and cars driving. The scene then shifts to an aerial view of a city with buildings
and a river in the background. The video then cuts to a red carpet event with
people walking and talking. The camera focuses on a man in a suit and a woman in a
dress. The video then shows a man in a suit and a woman in a dress walking down
the red carpet. The camera then focuses on a man in a suit and a woman in a dress
talking to each other. The video ends with the man in the suit smiling and walking
away.
```



Plain Text

```
The Suburban Palace Hotel, a notable landmark in Moscow, Russia. The hotel stands
tall, its white facade contrasting with the green sign that proudly displays its
name. The building itself is a blend of modern and traditional architecture, with
a flat roof and arched windows that add a touch of elegance. The perspective of
the photo is from street level, giving a sense of the hotel's imposing stature.
In the foreground, cars are parked, hinting at the bustling city life that
surrounds this serene structure. The sky above is a clear blue, dotted with a few
clouds, adding to the overall tranquility of the scene.
```