1 - Three engines (pick by what your video means)

```
Transcript-first . reads captions only . <$0.01 / 30-min . talking-head, lectures, podcasts.
Hybrid ......... transcript + sampled keyframes . ~$0.05-0.25 . webinars, tutorials, screen-share.
Native multimodal  watches frames + hears audio . ~$0.25-1.35 . demos, sports, surveillance.

Deciding question:  does the meaning live in the WORDS, or in the PICTURE?
```

2 - The five-stage pipeline (every engine shares it)

```
1 Acquire words ... existing captions (free) OR run speech-to-text (Whisper ~10% WER).
2 Clean & segment  stitch caption cues into sentences; chunk (~10k chars / 1k overlap).
3 Summarize ...... stuff / map-reduce / refine (see box 3).
4 Shape output ... bullets, chapter markers, JSON; attach timestamps (MM:SS).
5 Check .......... faithfulness gate - verify every claim against the source.
```

3 - Summarize a long transcript + the cost arithmetic

```
Map-reduce  parallel chunks then combine - fast; can't link distant sections.
Refine ..... running summary carried forward - reasons across; sequential, slower.
Stuffing ... whole transcript in one long-context prompt - default for one-shot summaries.

30-min video:  transcript ~6,000 tokens = <$0.01  |  multimodal 1 FPS ~300 tok/sec
  -> 1,800s x 300 = 540,000 tokens ~ $1.35 (full-res)  |  low-res ~100 tok/sec ~ $0.23.
```

4 - Build vs buy (rent the model, own the pipeline)

```
Rent APIs ... frontier model + hosted STT - fastest, pay per use; wins at low/spiky volume.
Self-host ... Whisper + open LLM - fixed cost; wins at high steady volume or strict privacy.
```

5 - Ship gate - quality + legal (context, not legal advice)

- [ ] Pick the engine by content, not habit: use frames only when the meaning is visual.
- [ ] Ground every claim and every timestamp to the transcript line that supports it - let users click to verify.
- [ ] Score faithfulness with an LLM-as-judge, not word-overlap (ROUGE) - hallucination is the dangerous error.
- [ ] Cache long videos you re-query; compute the rent-vs-self-host crossover before committing.
- [ ] Access captions through sanctioned API paths only - most free transcript scrapers forbid commercial use.