Generative Video: Vendor Selection & Integration Checklist

1 - Choose the camp BEFORE the model

- [ ] Does the content have to stay on our own servers (medical, legal, unreleased footage)? If yes -> self-host open weights.
- [ ] Is volume high and sustained enough that fixed GPU cost beats per-clip rental? If yes -> open weights may win past the crossover.
- [ ] If both answers are no (true for most products starting out) -> rent a closed API and move on to building the product.

2 - Shortlist by hard constraints first, rank by quality second

- [ ] Rule OUT by: data residency, max latency, max price per clip, required resolution and clip length.
- [ ] Only THEN rank survivors by quality (Artificial Analysis Video Arena Elo). The popular, the best, and the cheapest are usually 3 different models.
- [ ] Match the use case: edit existing footage -> Runway (Aleph V2V). Pro color pipeline / HDR -> Luma Ray3. Already on Google Cloud -> Veo 3.1. Top blind-vote quality -> Kling 3.0. Best value -> Hailuo 02. High-volume drafts -> Pika 2.5. Brand + coherent dialogue -> Sora 2.

3 - Build the async pipeline on day one

```
Pipeline: Submit -> [MODERATE prompt] -> Queue (job ID) -> Generate (poll 10-20s OR webhook)
          -> [MODERATE output] -> Label (C2PA + watermark) -> Store + deliver.
Gate 1: moderate the prompt BEFORE dispatch - once submitted, you have paid for the job.
Gate 2: preserve provenance labels through every re-encode/edit - do not strip them.
Put a thin internal layer in front of the vendor: fail over / swap a sunsetting model, no rewrite.
```

4 - Do the cost math once - and set a hard ceiling

- [ ] Per clip = seconds x $/sec. Monthly = clips x per-clip. Real monthly = x (generations per kept clip) - usually about 3.
- [ ] Worked example: 8s x $0.10/sec x 10,000 clips = $8,000/mo -> x3 retries = $24,000/mo. Plan for the x3 number.
- [ ] Biggest levers: model tier (budget at $0.03/sec cuts ~2/3) and retries (cheap draft first, premium only on final render).
- [ ] Set a hard spend ceiling and a max-cycles-per-job cap so a runaway feature cannot blow the launch budget.

5 - Compliance & licensing - clear before launch

- [ ] EU AI Act Article 50 (in force 2 Aug 2026): mark + disclose AI-generated video. C2PA is an accepted marking method.
- [ ] Closed models (Sora 2, Veo) already sign output by default - your job is to PRESERVE the label to the viewer.
- [ ] Open-weight licenses differ: Apache 2.0 (Wan, LTX-Video, CogVideoX) = full commercial use; HunyuanVideo excludes EU/UK and gates large deployers. Read it before you build.