AI Video Cost Model Worksheet

Plug your workload into the per-hour-of-video cost model. May 2026 unit prices.

1. Closed-API unit prices (USD per 1M tokens, except where noted)

Model	Input $/1M	Output $/1M	Cache read	Best for
Gemini 2.5 Flash-Lite	0.10	0.40	0.01	Classification, routing
Gemini 2.5 Flash	0.30	2.50	0.03	Default video workload
Gemini 2.5 Pro	1.25	10.00	0.125	Long-context video
GPT-5	0.625	5.00	varies	Workhorse reasoning
GPT-5.5	5.00	30.00	varies	Flagship reasoning
Claude Haiku 4.5	1.00	5.00	0.10	Fast structured output
Claude Sonnet 4.6	3.00	15.00	0.30	Long-form, agents
Claude Opus 4.7	5.00	25.00	0.50	Hardest reasoning
OpenAI Realtime audio	32	64	0.40	Realtime voice
Gemini Live audio	3.00	12.00	varies	Realtime voice/video

2. Audio AI and generative video — cost per hour of content

Audio (ASR) per hour processed
Deepgram Nova-3 batch: $0.26
Deepgram Nova-3 streaming: $0.46
OpenAI gpt-4o-mini-transcribe: $0.18
OpenAI gpt-4o-transcribe: $0.36
AssemblyAI Universal Streaming: $0.15
Azure Speech real-time: $1.00
Google STT V2 base: $0.96
Self-host Whisper (compute only): ~$0.03

Generative video per hour of output
Hailuo 02 Standard 768p: $162
Runway Gen-4 Turbo: $180
Veo 3.1 Lite 720p: $180
Kling 3.0 720p: $270
Sora 2 Standard 720p: $360
Sora 2 Pro 720p: $1,080
Sora 2 Pro 1024p: $1,800
Veo 3 with native audio: $2,700

3. Worked example — 100,000 MAU video conferencing product

Workload (cheap-path build)	Volume / month	Monthly cost
Live captions (AssemblyAI Universal Streaming)	800,000 hr	$120,000
Real-time translation (Gemini Live, 20% of calls)	160,000 hr	$172,800
AI meeting note-taker (Gemini 2.5 Flash)	800,000 calls	$5,120
Weekly digest summaries (Gemini 2.5 Flash)	400,000 digests	$8,450
Vector storage (pgvector self-hosted)	1.8B embeddings	$2,000
Compliance + ops overhead (1.5× multiplier)	—	~$45,000
Total cheap-path build	—	~$353,000
(Naive build for comparison)	—	~$700,000

4. Cost levers — stack them all

Route to cheaper tier (5x to 30x) · Prompt caching (70% to 90% off cached) · Batch (50% off) · Low-resolution media (3x on video) · Context truncation (20% to 60%) · System-prompt compression (10% to 30%) · Self-host on open weights (5x to 20x) · Edge inference (up to 100x) · Async pipelines (50% to 80%) · Volume contracts (20% to 40% above $50K/mo) · Cost-per-request monitoring (10% to 30%) · Spot GPUs (50% to 80%).