RAY SUMMIT 2026 RECAP · TALK 03

Beyond Text-to-Video

Omni-Context Generation in a 33B Model

Native multimodal isn’t free — what MiniMax H3 chose, what it costs, and how the the community made it fast



Sattaya Singkul (Jo)

Co-founder · vLLM Bangkok

Specialist Lead of AI Solution & Platform · True Digital Group

24 OCT 2026 · ASIC ROOM, MS SIAM TOWER · vLLM BANGKOK × AGICAFET

Speaker

Sattaya Singkul

Current

truedigital — **Specialist Lead of AI Solution and Platform**

AI Technical

KBTG — **Research Engineer**

synthesia — **Research Scientist (UK - contract)**

DOLBY — **Speech/Audio Researcher (AUS - contract) contract)**

NECTEC a member of NSTDA — **Research Assistant (NCR)**

Business and Technical

THAILAND MENTAL HEALTH — **Product Owner (Mental Health Tech)**

AI Community

vLLM BANGKOK — **Co-founder, vLLM Bangkok**

Speech AI
Embedded AI
LLM AI
AI Research
Human Computer Interaction
AI Solution

Award

CAT — **Winer, CAT Cyfense Cybercorp Contest 2016**

Bronze Medal, SuperAI Engineer SS1

Education

KMITL พระจอมเกล้าลาดกระบัง

M.Sc. Information Science (innovator's promising with honor scholarship) scholarship)

B.Sc. Information Technology (first class honor) honor)

11 AI Research Papers

Research — CellPress, Springer, acm, MDPI, mit media lab, IEEE

Collaboration

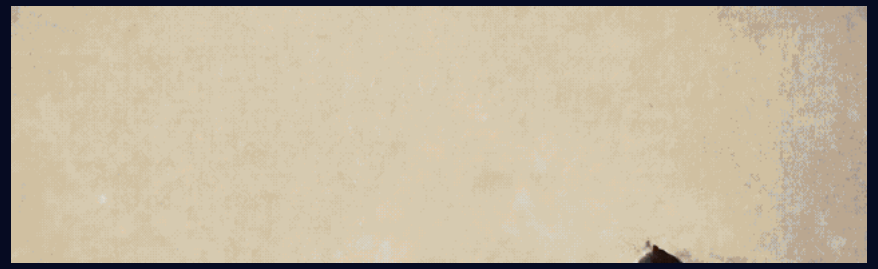
Google, tuya, CHUBB, ARKInsights

AI Committee

SuperAI Engineer, Springer, IEEE, acm

THE HOOK — FROM MINIMAX'S TALK

H3 is #1 in video editing. Why does it win — and what does it cost?



#1 · Elo 1,132

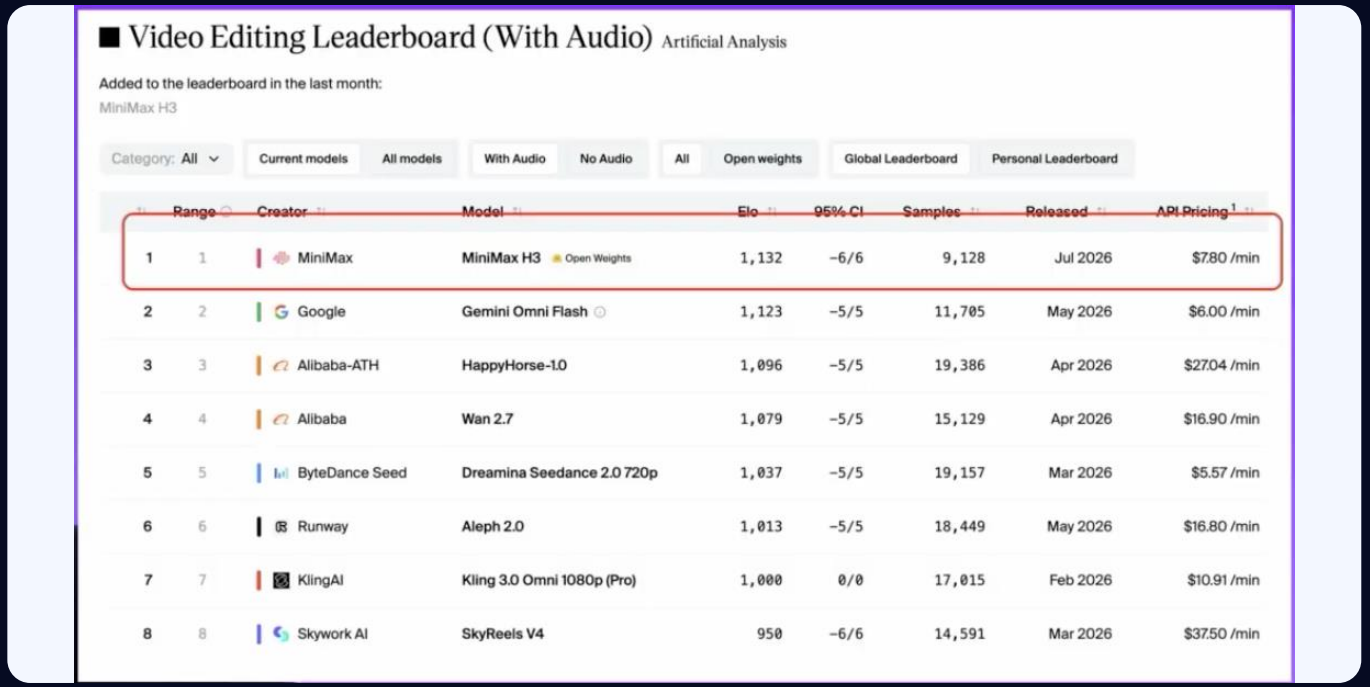
Video Editing with Audio · Artificial Analysis

#1 · 1,390

Arena Video Edit

Open weights

33B · download and self-host



■ Video Editing Leaderboard (With Audio) Artificial Analysis

Added to the leaderboard in the last month:
MiniMax H3

Category: All | Current models | All models | With Audio | No Audio | All | Open weights | Global Leaderboard | Personal Leaderboard

Rank	Range	Creator	Model	Elo	95% CI	Samples	Released	API Pricing[1]
1	1	MiniMax	MiniMax H3 Open Weights	1,132	-6/6	9,128	Jul 2026	$7.80 /min
2	2	Google	Gemini Omni Flash	1,123	-5/5	11,705	May 2026	$6.00 /min
3	3	Alibaba-ATH	HappyHorse-1.0	1,096	-5/5	19,386	Apr 2026	$27.04 /min
4	4	Alibaba	Wan 2.7	1,079	-5/5	15,129	Apr 2026	$16.90 /min
5	5	ByteDance Seed	Dreamina Seedance 2.0 720p	1,037	-5/5	19,157	Mar 2026	$5.57 /min
6	6	Runway	Aleph 2.0	1,013	-5/5	18,449	May 2026	$16.80 /min
7	7	KlingAI	Kling 3.0 Omni 1080p (Pro)	1,000	0/0	17,015	Feb 2026	$10.91 /min
8	8	Skywork AI	SkyReels V4	950	-6/6	14,591	Mar 2026	$37.50 /min

minimalist-product-ad-generator



music-video-subtitle-generator



3d-animation-short-generator



co-op-game-intro-generator



papercraft-stop-motion-explainer



paper-collage-explainer-generator



brand-promo-video-generator



handdrawn-live-video-generator



How to do that?

MiniMax' s bet: multimodality should be native, not assembled

1 · Preserve core intelligence
New modalities must not weaken language or reasoning

2 · Integrate across modalities modalities
Text and visual tokens interact — not stay separated inside

3 · Model space and time
Relationships across frames, not single single images

4 · Scale to real inputs
Long videos, many references, interaction history

ASSEMBLED

T2V → TTS → Lip-sync → Upscale

Every hop drops context → audio drifts, identity breaks

NATIVE

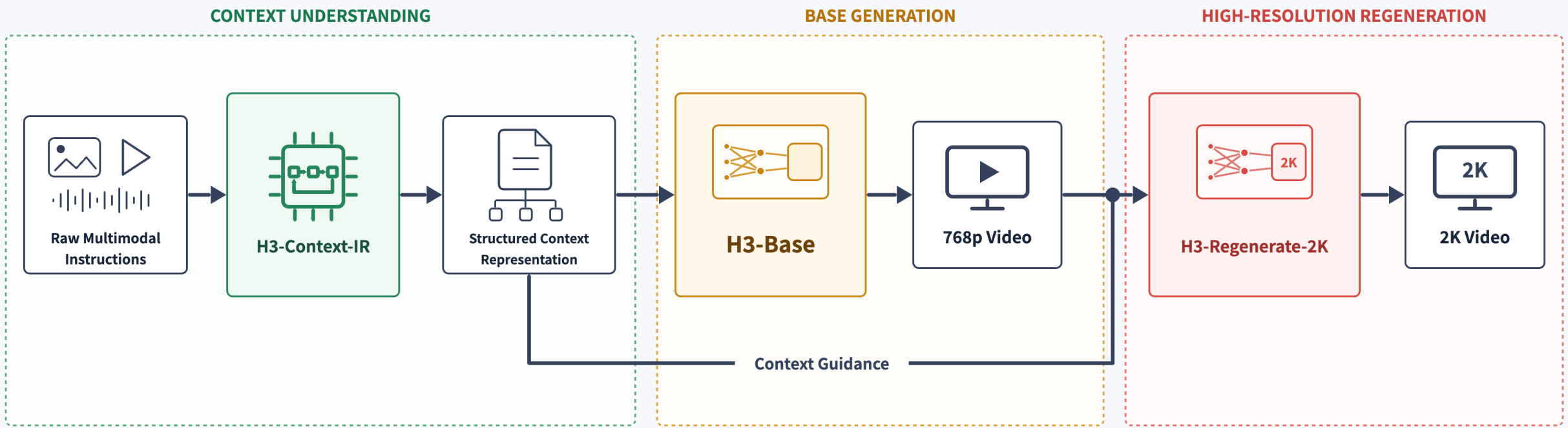
Text + image + video + audio → one model

Every modality sees every other, in every layer → sync emerges

Answer ① → H3 wins because it is native — audio and video share every layer. But native has a cost. → Act 2

What is H3?

MiniMax H3 System Overview



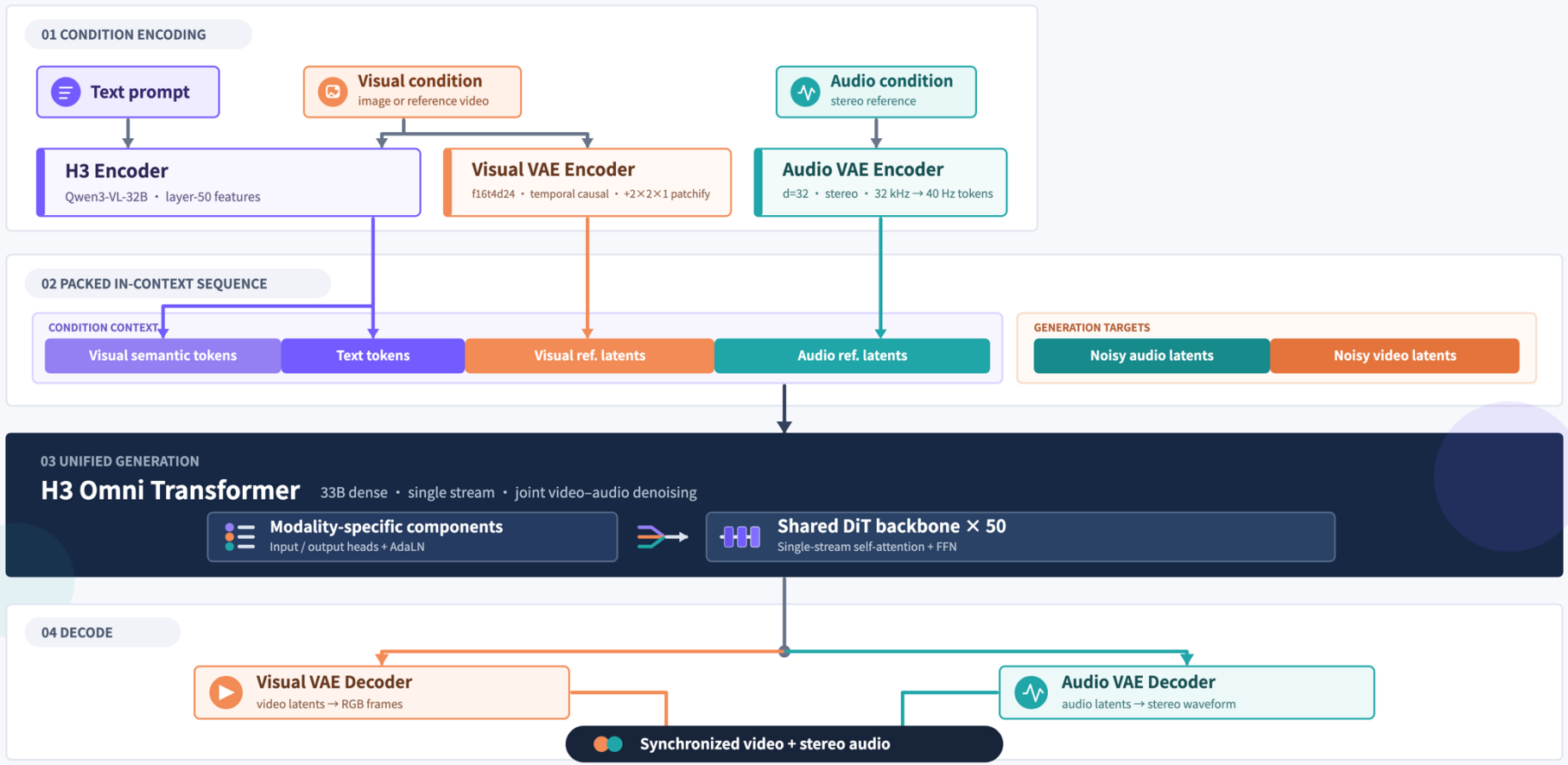
H3-Context-IR is a hosted preprocessing and orchestration system designed for free-form multimodal inputs. (private)
H3-Base: Generates audio and video based on the H3-Context-IR output, producing results at 768p resolution. (public)
H3-Regenerate-2K: Feeds the 768p result together with the original context back into H3 to regenerate the output at 2K resolution. (private)

ACT 1 · THE CHOICE

-

H3 Architecture

What is H3?



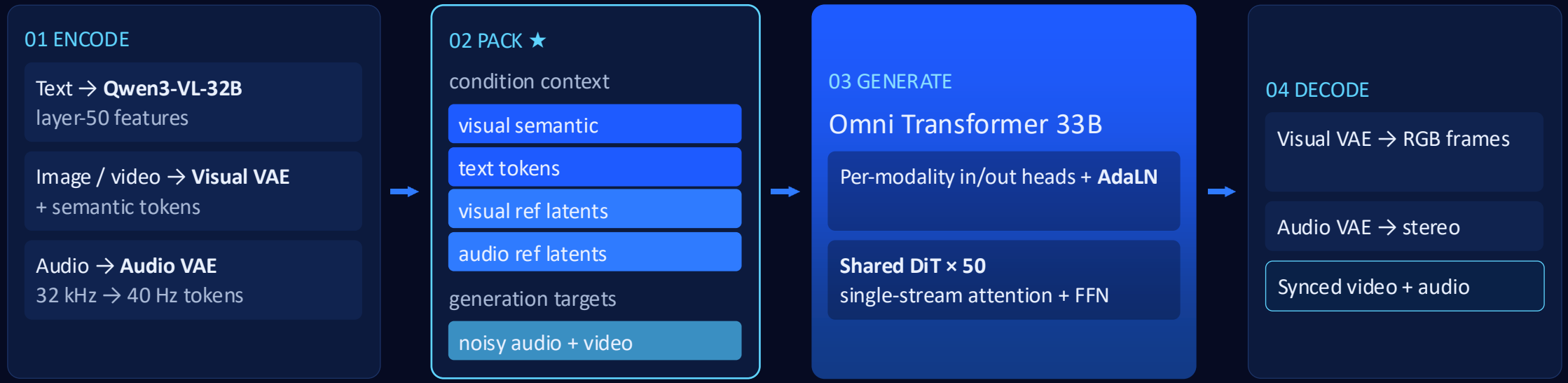


Is it cost and thread-off?



Inside H3-Base: 4 stages, one sequence

Full system: Context-IR (API) → H3-Base (open weights · this slide) → Regenerate-2K (API) — one context flows through every stage



Key → ★ Stage 02 is what “native” means: one sequence for everything. Sync comes for free — but the sequence gets very long.

ACT 2 · THE COST — OUR BREAKDOWN

A 10 s clip takes 82 s on 8 GPUs — here's why

COST BREAKDOWN · H3 · 10 s clip · 1344×768

① Memory	**~128 GB**
② Sequence length	**58,758 tokens**
③ Latency	**49 denoise steps**
④ Concurrency	**× every user**
RESULT	**82.2 s · 8×B300**

① **Memory** — does not fit one 80 GB GPU

② **Sequence** — attention is $O(n^2)$: ~3.46B pairs per head, per layer per layer

③ **Latency** — ~8× slower than playback

④ **Concurrency** — minutes per job; many users need a queue queue

Next → Act 3 fixes the 4 bottlenecks one by one — and shows why each fix works

How to fix it?

ACT 3 · FIX ① MEMORY — OUR DEEP DIVE

128 GB → 16 GB cards: precompute + quantize

AdaLN MLP
~13B params

→

Lookup table
~161 MB *

Why it works: AdaLN only sees the timestep (~50 values) and the modality → a finite set of outputs. Compute once, identical result.

BF16 checkpoint · 123.6 GB

INT8 · ~67 GB

Pruned INT8 + NVFP4 · ~42.5 GB

RESULT · COMMUNITY, FIRST 48 H

Consumer 8–16 GB
RTX 5080: 5 s clip + audio in ~3.5 min

Workstation 4090 / 5090
5090 + sol-attn: 5 s clip in 50 s

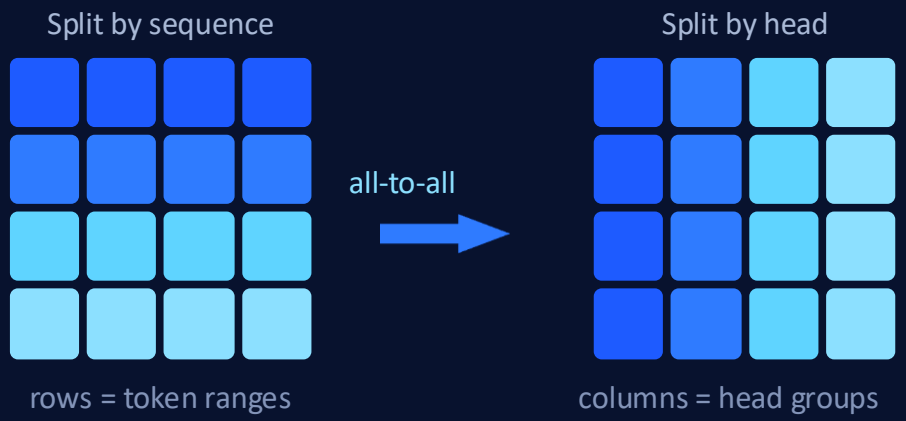
Datacenter 4×B300 / 8×GB200
8×GB200: 124 frames, 50 steps in 6.9 s

Apple MLX · DGX Spark
Spark, 12 s clip: 480p 15 min → 1080p 2.5 h

Reuse → Find large weights driven by a small, discrete input — and precompute them. e.g. TTS with fixed voice presets

ACT 3 · FIX ② SEQUENCE LENGTH — OUR DEEP DIVE

59K tokens → Ulysses splits attention across GPUs



Color = GPU 0–3 · after attention, all-to-all back to the left layout

Why: attention needs the full sequence for each head. Swapping the split axis twice axis twice per layer moves far less data than shipping all K/V.

Constraint: GPU count must divide 56 heads (2, 4, 7, 8, 14…)

+ Layerwise offload

Keep r blocks on GPU, prefetch the rest from host RAM. r = 35 → HBM −37.5% for +5.1% latency

+ Disaggregated encoder

Encoder runs once per request, DiT runs 49× → separate stages, scaled independently

→ Fix ④ builds on this

Independent stages are exactly what Ray orchestrates across many requests

ACT 3 · FIX ③a LATENCY — FROM MINIMAX'S TALK

Diffusion-native runtime: 82.2 → 56.9 s, lossless

	LLM serving	Diffusion serving (H3)
Loop	Decode token by token, unknown length	Fixed number of denoise steps (49)
Memory	KV cache grows with length	Full 59K sequence, every step
Output	Text stream	MP4 file, video + audio

WHAT vLLM-OMNI ADDS

- Task routing: T2VA / FL2VA / Ref2VA
- Shared loading of encoder + VAEs
- Parallelism + offload for ~128 GB
- OpenAI-compatible API

Diffusers 82.2 s → vLLM-Omni 56.9 s · 1.445×, lossless

8×B300 · 50 steps · same prompt & seed



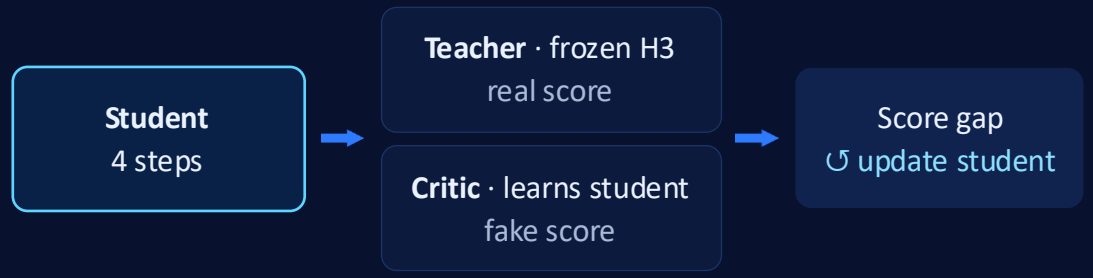
DiT 51.8 s (91%) · VAE 1.0 · transport ~2.5 · MP4 1.5 — strip padding, fuse kernels, convert frames to uint8 on GPU

Lesson → Profile every stage first. DiT is 91% — the biggest lever. Shrink it, and the other 9% becomes the next bottleneck (Amdahl).

ACT 3 · FIX ③b FEWER STEPS — OUR DEEP DIVE

FastH3: 49 → 4 forwards, faster than playback

DMD2 TRAINING LOOP



Why 4 steps is enough: the student learns to match the teacher's output distribution — not to retrace its 49-step path.

+ VSA: keep ~10% of video↔video attention tiles. Text and audio stay dense stay dense — few tokens, but they carry meaning and sync.

10 s CLIP · 8×B300 · COMPLETE MP4

Playback · 10.1 s

FastH3 · 8.7 s (real-time factor 0.86)

DiT 5.5 · VAE 1.25 · I/O + MP4 1.75 s

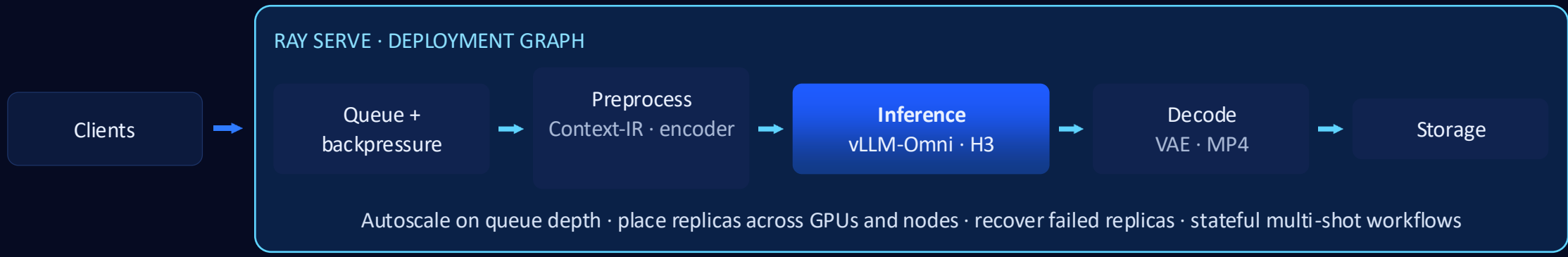
Non-DiT share is now **34%** — Amdahl, as predicted

1×B200, 15 s clip: 678.7 → 47.2 s = **14.4×**

v1 limits: text → video + audio only · fused at load, no per-request switching

ACT 3 · FIX ④ CONCURRENCY — FROM MINIMAX'S TALK

Many users: Ray orchestrates H3 in production



vLLM-Omni

Makes **one request** fast — parallelism, kernels, fewer steps

Ray

Keeps **many requests** alive — queues, scaling, recovery

Lesson → Video jobs take seconds to minutes, not milliseconds. Design async jobs + queues from day one — don't let timeouts teach you.

ACT 4 · BEFORE → AFTER

From 8× slower than playback → faster than playback

Latency · 10 s clip, 8×B300	82.2 s →	**8.7 s**
Checkpoint size	123.6 GB →	**~42.5 GB**
Minimum hardware	8 GPUs →	**16 GB card**

82.2 s = base H3 (Diffusers, 50 steps) · 8.7 s = FastH3 (4 steps) — different test runs, so read it as the scale of change, not a matched benchmark. Quality: still #1 in video editing.
#1 in video editing.

[New!!] https://huggingface.co/FastVideo/FastH3-Omni-Comfy

ACT 4 · TAKE IT HOME

Next new model? Ask the 3 questions we asked H3

① The choice? · Act 1 → ② The cost? · Act 2 → ③ The fixes? · Act 3

Question	H3's answer	Ask of your own system
①Choice	Native — every modality in one sequence	Native or assembled? How many hops?
②Cost	128 GB · 59K tokens · 49 steps · 82 s	Measure memory, tokens per request, forwards, end-to-end time
③Fixes	Precompute + quantize · Ulysses · diffusion runtime + distillation · Ray	Fix the biggest bottleneck first — and label every fix lossless or lossy

Notice → These 3 questions were the structure of the last 20 minutes. They work for video, TTS, STT or LLMs.

REFERENCES

Every number in this deck, traced to its source

[1] MiniMax — H3 session, Ray Summit 2026 (on-stage slides)
vllm.ai/events/vllm-conference/2026

[2] MiniMax — H3 launch blog (31 Jul 2026)
minimax.io/blog/minimax-h3

[3] MiniMax H3 model card + Community License
huggingface.co/MiniMaxAI/MiniMax-H3

[4] vLLM-Omni Team — H3 production serving (1 Sep 2026)
vllm.ai/blog/2026-09-01-minimax-h3-...

[5] vLLM Recipes — MiniMax-H3
recipes.vllm.ai/MiniMaxAI/MiniMax-H3

[6] FastVideo — FastH3 Preview v1 (27 Aug 2026)
haoailab.com/blogs/fasth3-preview

[7] ComfyUI — H3 Day-0 support
blog.comfy.org/minimax-h3-day-0-...

[8] Spheron — Deploy MiniMax H3 on GPU cloud
spheron.network/blog/deploy-minimax-h3-...

[9] Runpod — H3 and what it takes to run it
runpod.io/blog/minimax-h3-...

[10] Artificial Analysis — Video Editing Leaderboard
artificialanalysis.ai · Arena: arena.ai

[11] The Batch — H3 is only minimally open
deeplearning.ai/the-batch/...

[12] Yin et al. — DMD2, arXiv:2405.14867 (2024)

[13] Zhang et al. — VSA, arXiv:2505.13389 (2025)

[14] Jacobs et al. — DeepSpeed-Ulysses, arXiv:2309.14509 (2023)

[15] Peebles & Xie — DiT / AdaLN-Zero, arXiv:2212.09748 (2023)

TAKE-HOME

Native isn’t free — but every cost can be engineered down

① What’s the choice?

② What does it cost?

③ How is it fixed?

Try this week: measure your system’s 4 numbers — memory · tokens per request · forwards · end-to-end time — then fix the biggest one first.

Backup

AR LLM vs Diffusion LLM

