Leoma

Decentralized Incentive Network for
AI Video Generation Model Development

Subnet 99 | Bittensor | Version 1.0 - April 2026

1. Abstract

Sora cost OpenAI $15 million a day in compute and still shut down. The video generation market is projected to reach $2.2B-$4.1B by 2030, but no centralized provider has found unit economics that work. Meanwhile, open-source models running on consumer GPUs are closing the quality gap fast.

Leoma is a Bittensor subnet that turns this into an opportunity. Independent developers compete to build AI video generation models. Independent validators evaluate them across six quality dimensions using structured GPT-4o assessment, stake-weighted consensus, and a 100-task rolling window. The top-ranked model earns 100% of TAO emissions. Everyone else earns nothing.

The result: a system where the best available video generation model is always the one serving users, verified by people with economic skin in the game, not by the team that built it.

2. Problem

AI video generation has gone from low-resolution curiosities to photorealistic productions with coherent motion, plausible physics, and cinematic composition - all since 2023. But the ecosystem producing these capabilities has structural problems that limit who benefits and whether the economics can hold.

Unsustainable Economics

Video generation requires 10-100x more compute than image generation. A single 10-second 1080p clip costs $0.10-$2.00 in inference alone. Sora is the clearest example of what happens when you ignore this: massive consumer interest, $15M/day in compute, no viable revenue path. Every company in this space faces the same mismatch between centralized GPU inference costs and what users will actually pay.

Concentration of Control

Google (Veo), Runway (Gen-4.5), Kuaishou (Kling), and Pika control access to the most capable models. They set pricing, availability, content policies, and terms of service unilaterally. Users are locked into platform-dependent workflows with no portability. If a provider raises prices, restricts access, or shuts down, users have no recourse. Sora users learned this the hard way.

Opaque Quality Standards

There is no standardized way to compare model quality across providers. Marketing benchmarks are cherry-picked. "Cinematic quality" and "photorealistic output" are claims nobody can verify. Without independent evaluation, providers face no external pressure to improve beyond "good enough."

Innovation Bottleneck

Independent researchers and small teams build impressive video generation models but have no path from innovation to economic reward. Academic papers demonstrate breakthroughs. The gap between a research prototype and a production deployment stays vast. Talent and compute flow to well-funded companies, not to the best ideas.

3. Market Context

The AI video generation market sits at $700M-$950M in 2025, projected to reach $2.2B-$4.1B by 2030 at 18-25% CAGR (Fortune Business Insights, Grand View Research). Several dynamics make this moment specific:

- **Cost reckoning.** The industry is shifting from "demonstrate capability at any cost" to "build sustainable unit economics." Sora accelerated this.
- **Open-source acceleration.** Wan 2.1 (Alibaba, Apache 2.0), HunyuanVideo (Tencent), and LTX-Video (Lightricks) run on consumer GPUs with 14-24GB VRAM. Wan 2.1 hit 2.2M+ downloads on Hugging Face within weeks of release.
- **Enterprise adoption.** Marketing, advertising, and corporate communications account for 34% of AI video usage. Synthesia alone exceeds $100M ARR.
- **Quality convergence.** The gap between open-source and proprietary models went from "unusable vs. impressive" to "competitive vs. leading" in about 12 months.

The leading proprietary models (Runway Gen-4.5, Google Veo 3.1, Kling 2.6) produce 4K video up to 30 seconds. Leading open-source models (Wan 2.1, HunyuanVideo, CogVideoX) produce 720p-1080p up to 10 seconds. The remaining gap in resolution, duration, multi-character consistency, and controllability is the opportunity Leoma targets.

Why now and not three years ago:

Three years ago, none of this was possible. Open-source video models did not exist at usable quality, GPU costs made decentralized inference a non-starter, and Bittensor was too early to support something this complex. All three things changed in a short window. Open-source models now run on consumer hardware. Decentralized GPU networks have pushed inference costs down hard. Bittensor has 128+ active subnets and an economic model that works. The gap between open-source and proprietary video quality is small enough to close, the infrastructure is ready, and nobody has built the incentive layer to coordinate it. That window will not stay open forever.

4. Solution

4.1 Architecture

Leoma is a three-tier system. The Bittensor blockchain stores miner commitments, validator weight-settings, and manages TAO emissions and staking. Off-chain infrastructure handles the compute: task generation, video inference, quality evaluation, score aggregation, and data storage. This separation keeps the protocol fast while maintaining on-chain verifiability for the decisions that matter - weights and emissions.

4.2 Participants

Miners build or fine-tune TI2V models that accept a conditioning image and text prompt and produce video output. They upload model weights to Hugging Face with verifiable commit SHAs, deploy to Chutes (serverless GPU inference), and register on-chain by committing a JSON record containing model name, revision SHA, and Chute ID. Repository names must follow the convention {username}/leoma-{description}-{hotkey} to prevent impersonation.

Validators poll for evaluation tasks every 60 seconds, download task artifacts and miner-generated videos from object storage, score each output using GPT-4o across six quality dimensions, and

submit signed pass/fail evaluations with cryptographic hotkey signatures. They set on-chain weights each epoch (~36 minutes) based on the aggregated ranking.

The Owner API coordinates everything: generating evaluation tasks every 20 minutes from curated source videos, orchestrating miner inference via Chutes, aggregating validator evaluations using stake-weighted consensus, computing eligibility and rankings, and publishing scores via REST API for public audit.

4.3 Protocol Flow

Step 1 - Task Generation (every 20 minutes). The Owner API selects a 5-second, one-shot source clip with no scene changes from a curated pool. It extracts the first frame as a PNG conditioning image and uses GPT-4o to generate a ~70-word benchmark prompt describing scene composition, subject attributes, expected motion, and camera guidance.

Step 2 - Miner Inference. For each valid miner, the Owner API calls the miner's Chute endpoint with the first frame and prompt. The model generates a video, uploaded to object storage.

Step 3 - Validator Evaluation. Each validator independently downloads the task artifacts and every miner's video. For each video, the validator extracts 12 frames at 3fps and sends them alongside the conditioning frame and text prompt to GPT-4o. GPT-4o returns scores across six dimensions (0-100 each), a confidence score, and identified issues and strengths. The validator applies pass/fail logic and submits a signed batch to the API.

Step 4 - Scoring (every 30 minutes). A background task collects all validator evaluations for the current window (100 consecutive tasks). For each (task, miner) pair, individual votes are aggregated using stake-weighted consensus. The system checks eligibility (80%+ completeness in the window), ranks eligible miners using the dominance algorithm, and updates the ranking table.

Step 5 - Weight Setting (every epoch, ~36 minutes). Each validator fetches the current winner and sets on-chain weights: the winner gets 1.0, everyone else gets 0.0. If no eligible winner exists, weight goes to UID 0 - emissions burn rather than reward inadequate models.

5. Evaluation

The evaluation mechanism is the core of the system. Everything else exists to support it. If evaluation breaks, nothing else matters.

The failure mode is obvious. Weak evaluation means miners game it. Gamed evaluation means the wrong model wins. The wrong model serves users, users leave, and the network dies. So we built evaluation like it is the product. Because it is.

5.1 Scoring Dimensions

Dimension	Weight	Critical?	What It Measures
First Frame Fidelity	25%	Yes	How faithfully the video preserves the conditioning frame's content, composition, and detail
Prompt Adherence	25%	Yes	How accurately the video follows the text description for scene, subjects, actions, and attributes
Temporal Consistency	20%	Yes	Whether subjects, backgrounds, and lighting stay coherent across frames without flickering or drift
Motion Quality	15%	No	Naturalness, fluidity, and physical plausibility of motion
Visual Quality	10%	No	Resolution, sharpness, color accuracy, absence of artifacts
Camera Composition	5%	No	Adherence to camera movement instructions and cinematographic quality

First Frame Fidelity and Prompt Adherence get the highest weights because TI2V generation is fundamentally about following the user's inputs. Temporal Consistency is the hard problem that separates video from image generation. Motion Quality captures what makes video compelling. Visual Quality and Camera Composition matter, but a video that follows instructions with natural motion beats a sharp, well-composed video that ignores the prompt.

5.2 Pass/Fail Decision

Each validator makes a binary call per miner per task. PASS requires a weighted score of 75+ overall AND each critical dimension (First Frame Fidelity, Prompt Adherence, Temporal Consistency) scoring 50+. Everything else is a FAIL. The critical floors prevent a model from gaming the system by excelling only in non-critical areas. If GPT-4o returns an unparseable response, the score defaults to 0.

5.3 Stake-Weighted Consensus

No single validator determines a miner's fate. For each (task, miner) pair, the system collects all validator votes weighted by stake. If the ratio of stake-weighted passes to total stake exceeds 0.5, the miner passes. A single malicious validator cannot flip results unless they control more than 50% of total stake. Creating many low-stake Sybil validators cannot outweigh fewer high-stake honest ones.

5.4 Eligibility and Ranking

Miners must complete at least 80% of the last 100 consecutive tasks (~33 hours of evaluation) to be eligible. This prevents cherry-picking easy tasks, intermittent participation, and ghost registrations.

Eligible miners are ranked using a dominance algorithm. Miners are ordered by on-chain registration block (earliest first). A later miner can only displace an earlier one if their pass rate exceeds the incumbent's by more than 5%. This threshold prevents constant turnover from marginal improvements - a new miner scoring 80.1% does not immediately displace one at 80.0%. Between miners with similar pass rates, the earlier registrant holds position, which discourages copying an existing model with minor tweaks. If no miner clearly dominates, the system falls back to highest absolute pass count.

5.5 Winner-Take-All

The top-ranked miner receives 100% of emissions. Everyone else gets nothing. There is no reward for second place.

Most incentive systems split rewards across the top performers. That sounds fair. In practice, it kills innovation. If you earn 60% of emissions for a model that is 95% as good as the leader, why push harder? Proportional rewards create a comfortable middle where "good enough" pays well. Winner-take-all eliminates that middle entirely. You either build the best model or you keep working until you do. Second place earns nothing. That is the point.

Since only one model wins, miners invest in building something genuinely new rather than running many similar variants. And users always get served by whatever model currently holds the top rank.

When no eligible winner exists, emissions go to UID 0. They burn. The system does not reward inadequate work.

6. Anti-Manipulation

Plagiarism detection. SHA-256 hash of sorted model weight files. Identical models get flagged - only the earliest registrant (by block number) is valid. Duplicates are excluded from evaluation.

Revision pinning. Miners commit to specific Hugging Face commit SHAs on-chain, not branch names. The deployed revision must match exactly. No silent model swaps between evaluations.

Evaluation integrity. The GPT-4o prompt instructs the model to ignore embedded text, watermarks, and adversarial content in video frames. Scores are clamped to [0, 100]. Parse errors default to score=0 and fail.

Validator authentication. All evaluation submissions are signed with the validator's cryptographic hotkey. Manipulating consensus requires controlling a majority of economic stake - prohibitively expensive.

Naming enforcement. Hugging Face repositories must start with "leoma" and end with the miner's hotkey. No impersonation.

7. Evaluation in Context

The video generation research community has developed several benchmark suites. VBench (CVPR 2024 Highlight) evaluates 16 dimensions using specialized models and features 50+ models on its leaderboard. VBench-2.0 (March 2025) advances to 18 dimensions using Video Language Models. VideoScore (EMNLP 2024), trained on 37,600 human-annotated videos, achieves 77.1 Spearman correlation with human judgment - the strongest automated result, but still imperfect.

How Leoma differs:

Aspect	Industry Standard (VBench)	Leoma
Frequency	One-time benchmark run	Every 20 minutes
Task diversity	Fixed prompt set	Dynamic tasks from real video sources
Evaluator	Specialized models (DINO, CLIP, RAFT)	GPT-4o multi-modal (closer to human judgment)
Consensus	Single evaluation	Stake-weighted multi-validator consensus
Input modality	Text-to-video	Text+Image-to-video (tests conditioning fidelity)
Temporal coverage	Single evaluation point	100-task rolling window (~33 hours)

VBench answers "how does this model perform on a standard test set?" Leoma answers "is this model producing quality output across diverse, unpredictable inputs over time?" The first is useful for research. The second is necessary for production.

The Automated-Human Judgment Gap

Traditional metrics (FVD, CLIP-Score) show poor or negative correlation with human judgment in many settings. VBench gives near-identical scores to models with clearly different perceptual quality. GPT-4o achieves stronger correlation when given structured evaluation criteria rather than asked for a single holistic score.

Leoma addresses this four ways: structured prompts requesting scores across six specific dimensions rather than one quality score; multi-validator consensus averaging out individual biases; binary pass/fail aggregation reducing the impact of score-level inaccuracy (the system only needs to determine pass or fail, not precise ranking from continuous scores); and dimensions that map directly to what users care about - does it match the input, is the motion natural, is it visually coherent.

8. Revenue Model

Stream	Description
API Access	Pay-per-generation and subscription tiers, routed to the top-ranked model
Enterprise	Custom integrations with SLA-backed quality guarantees
TAO Emissions	Subnet participants earn emissions; quality attracts more stake via dTAO
Premium Features	Higher resolution, longer duration, priority queue, batch processing

Pricing ranges from a free developer tier (5-10 generations/day) through Creator ($15-30/month) and Professional ($50-150/month) to custom Enterprise and per-generation API pricing ($0.05-$0.50/generation). Decentralized GPU networks offer compute at significantly lower cost than centralized cloud providers, which gives Leoma room to price competitively while maintaining margins. These tiers are provisional and will be adjusted based on actual compute costs and market

response at launch.

Who pays:

A marketing agency generating 50 ad variations from a single product photo at a fraction of current production costs. A game studio prototyping cutscene animations from concept art before committing to full production. An e-commerce platform turning static product images into short video clips for social media at scale. A solo content creator who cannot afford Runway's pricing but needs video that looks professional.

9. Risks and Limitations

GPT-4o Dependency

The evaluation mechanism relies on OpenAI's GPT-4o. That is a centralized dependency inside an otherwise decentralized system. If OpenAI changes GPT-4o's behavior, pricing, or availability, evaluation is affected. The roadmap includes migration to open-source evaluation models (VideoScore variants, AIGV-Assessor) running on the Bittensor network. Phase 2 supplements GPT-4o with specialized models.

Evaluation Accuracy

GPT-4o's correlation with human judgment is strong for structured tasks but imperfect. Some quality distinctions obvious to humans will be missed. Binary pass/fail aggregation, multi-validator consensus, and the 100-task rolling window reduce the impact of individual errors. The system needs to get most evaluations approximately right, not every single one perfectly right.

Winner-Take-All Risk

Miners can invest heavily in model development and earn zero if another model is marginally better. That could discourage participation if the incumbent advantage becomes entrenched. The 5% dominance threshold creates a meaningful but not insurmountable barrier. New entrants know exactly what quality level they must exceed. The threshold can be adjusted via governance as the ecosystem matures.

Infrastructure Dependencies

Leoma depends on Hugging Face for model hosting and Chutes for serverless inference. Outages or policy changes from either could disrupt operations. Both are standard choices with large user bases and strong uptime records. The architecture allows migration to alternative providers if needed.

Security Threat Model

Threat	Mitigation
Model plagiarism	SHA-256 weight hashing; earliest registrant priority; duplicate invalidation
Evaluation manipulation	Multi-validator consensus; stake-weighted voting; structured GPT-4o prompts resistant to adversarial content
Validator collusion	>50% of total stake required to control consensus - economically prohibitive
Sybil attacks (miners)	On-chain registration requires TAO stake; plagiarism detection prevents duplication
Sybil attacks (validators)	Stake-weighted voting neutralizes low-stake Sybils
Task prediction	Source clips from diverse pool; prompts generated dynamically by GPT-4o
API impersonation	Hotkey signature verification on all authenticated endpoints
Evaluation gaming	Critical floors on three dimensions; clamped scores; consecutive window prevents selective participation

10. Roadmap

10.1 Where We Are Now

v0 - Miner Fine-Tuning Layer (Current)

Miners fine-tune Wan 2.2 (~14B parameters) using domain-specific datasets. The architecture is a latent diffusion / video transformer backbone with temporal attention. Outputs are evaluated via the multi-metric pipeline described in Section 5. Top-performing checkpoints already exceed the base model distribution in perceptual quality and coherence.

v1 - Production Model (Target: April 30 Launch)

The highest-scoring miner checkpoint, selected via validator consensus, becomes the basis for Leoma v1. We standardize the inference stack - sampling schedule, guidance scaling, temporal consistency settings - and deploy as a public API and product. Optimization targets:

- Inference latency vs quality tradeoff
- Stability under diverse prompt distributions
- Cost-efficient generation (throughput per GPU)

v1.x - Iterative Refinement (Post-Launch)

Miners continue submitting checkpoints. The best replace the serving model on a rolling basis. In parallel:

- Dataset expansion: high-quality video-text pairs, motion-rich and edge-case scenarios
- Evaluation expansion: adversarial prompts, long-horizon temporal consistency tests
- Introduction of partial model merging and distillation across top miners

10.2 The Leoma-Native Base Model: Why 25B+ Matters

Everything described above - v0 through v1.x - is built on top of someone else's model. Miners fine-tune Wan 2.2. They improve it, they specialize it, they push it beyond its original distribution. But the foundation is still Alibaba's architecture. The long-term vision for Leoma is to replace that dependency entirely with a model we build, train, and own.

That model is v2: a 25B+ parameter video generation architecture built from scratch. To understand why this matters, consider the current landscape. The largest open-source video generation model is approximately 14B parameters. No open-source project has built beyond that. In the language model space, Templar's achievement of training a 72B parameter model on Bittensor was a landmark event for the network - proof that decentralized coordination could produce models at a scale previously reserved for well-funded labs. Building a 25B+ video model is the equivalent challenge for video generation, and arguably harder.

Video models are not language models with pixels. They must maintain temporal coherence across frames - objects need to persist, lighting needs to stay consistent, motion needs to follow physics. A language model predicts the next token. A video model predicts the next frame while keeping every previous frame consistent. The computational and architectural challenges compound with every additional parameter. Scaling from 14B to 25B+ is not a modest increase in difficulty. It is a generational leap in architecture design, training infrastructure, dataset curation, and optimization.

A single team attempting this would need tens of millions in compute budget. Bittensor changes the economics. The incentive mechanism distributes the cost of model development across miners who are economically motivated to contribute. The evaluation framework ensures that only genuine improvements earn rewards. We are not building on Bittensor because decentralization sounds good in a whitepaper. We are building on it because there is no other way to fund a model this large without being Google or OpenAI.

The v2 architecture:

- Hybrid diffusion + transformer design
- Extended temporal context window for longer, more coherent sequences
- Multi-scale spatial-temporal attention

The training stack:

- Large-scale curated dataset across multiple video domains
- Contrastive alignment (CLIP-style) for text-video correspondence
- Temporal consistency regularization using optical flow and latent warping losses

We are not trying to be slightly better than existing open-source baselines. The goal is to surpass them in motion realism, temporal coherence, and controllability - and to prove that decentralized competition can get there faster than centralized capital.

10.3 After v2

v3 - Open Release + Subnet Integration

We publish the model weights and a technical paper. The Leoma-native architecture becomes the new training prior for miners. Instead of fine-tuning Wan or another external model, miners fine-tune on top of Leoma's own architecture. This unifies the entire loop: base model, miner specialization, evaluation, reintegration. Every improvement feeds back into the foundation.

End State

No dependency on external open-source models. A vertically integrated pipeline from data through base model through miner fine-tuning through evaluation through product. Competitive performance against closed AI systems - Runway, Kling, Veo - with lower cost and decentralized scaling.

11. Conclusion

Centralized video generation does not have sustainable unit economics. Sora proved that. Open-source models are closing the quality gap. Enterprise demand is growing. The opening is here.

Leoma fills it with a specific mechanism: miners compete to build AI video generation models, validators evaluate them across six dimensions with stake-weighted consensus over a rolling 100-task window, and the winner takes all emissions. The evaluation is the product. Get the rankings right and everything else follows - better models attract users, users generate revenue, revenue attracts better miners.

The hard part is not the architecture. It is whether we can attract developers good enough to build video models that compete with products backed by the best-funded companies in history, and whether our evaluation can stay accurate as those models improve. Phase 1 is live. We will find out.

References

Academic Benchmarks and Evaluation

Huang, Z. et al. "VBench: Comprehensive Benchmark Suite for Video Generative Models." CVPR 2024 Highlight.

Huang, Z. et al. "VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness." arXiv:2503.21755, March 2025.

He, X. et al. "VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation." EMNLP 2024.

Yuan, S. et al. "ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation." NeurIPS 2024 Spotlight.

Liu, Y. et al. "EvalCrafter: Benchmarking and Evaluating Large Video Generation Models." CVPR 2024.

AIGV-Assessor. "Benchmarking and Improving AIGV Quality Assessment." CVPR 2025.

Ge, S. et al. "On the Content Bias in Frechet Video Distance." CVPR 2024.

Unterthiner, T. et al. "Towards Accurate Generative Models of Video: A New Metric & Challenges." arXiv:1812.01717.

Liu, J. et al. "Frechet Video Motion Distance: A Metric for Evaluating Motion Consistency in Videos." ICML 2024.

Luo, G. Y. et al. "Beyond FVD: An Enhanced Evaluation Metrics for Video Generation Distribution Quality." ICLR 2025.

Qi, Z. et al. "Towards Holistic Visual Quality Assessment of AI-Generated Videos." CVPR Workshop 2025.

Fu, C. et al. "Video-MME: Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis." CVPR 2025.

Chen, D. et al. "MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark." ICML 2024.

Video Generation Models

Wan-Video Team. "Wan: Open and Advanced Large-Scale Video Generative Models." arXiv:2503.20314, March 2025.

Kong, W. et al. "HunyuanVideo: A Systematic Framework For Large Video Generative Models." Tencent, arXiv:2412.03603, December 2024.

Bittensor Protocol

Bittensor Foundation. "Bittensor Whitepaper." bittensor.com/whitepaper

Bittensor Foundation. "Dynamic TAO (dTAO) Whitepaper." bittensor.com/dtao-whitepaper

Bittensor Documentation. docs.learnbittensor.org

Market Research

Fortune Business Insights. "AI Video Generator Market Size, Share & Industry Analysis." 2025.
Grand View Research. "AI Video Generator Market Report." 2025.
"Why OpenAI Really Shut Down Sora." TechCrunch, March 2026.
Synthesia. "Synthesia Surpasses $100 Million in Annual Recurring Revenue." April 2025.

Decentralized AI

Aethir. "The $7 Trillion AI Arms Race: How Decentralized GPU Networks Offer a Smarter Path Forward." 2025.
io.net. "How Decentralized GPU Networks Are Powering the Next Generation of AI." 2025.

Disclaimer: This whitepaper is for informational purposes only. It does not constitute financial advice, an investment solicitation, or a guarantee of future performance. TAO emissions, subnet rankings, and market conditions are subject to change. Participation involves risks including technical failures, market volatility, and regulatory uncertainty.