YOUR VIDEO GENERATOR SECRETLY CONTAINS PHYSICS EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation models can produce high-quality, realistic videos, yet they often violate basic physical laws. Existing work largely treats these failures as evidence of missing capabilities that must be added through physical supervision, simulators, or additional training. In this paper, we find that video generators capable of producing physically better videos already exist within pretrained video models. We can find those physics experts by applying random masks in the video models. Therefore, the physical accuracy of the video generators can be improved without training. Across two model families, we also observe that at a fixed masking rate, these physics experts become denser as the model scale increases. In practice, they can be selected without ground-truth videos, reused across scenes, and obtained with comparable visual quality. Our results offer a new perspective on physical failures in video generation: physical capabilities may not be absent from pretrained models but rather distributed across the diverse subnetworks they contain.

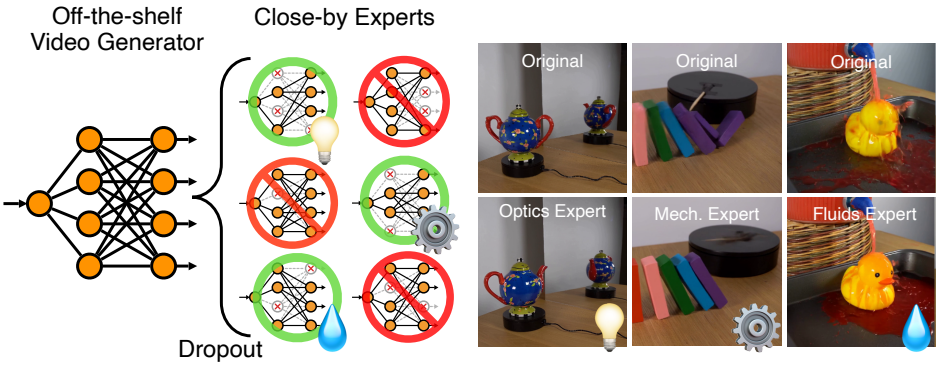

Figure 1: **Physics experts are around the pretrained video models.** We find that the subnetworks in the pretrained video generators are good at different physical domains, and they can be discovered easily by masking the model weights without training. For example, certain subnetworks can be better than the base model in Optics, Mechanics, or Fluid Dynamics.

1 INTRODUCTION

Recent advances in video generation models have shown remarkable quality of the generated videos, yet these models often fail at fundamental physical understanding. Even in highly realistic scenes, solid objects may seamlessly pass through fabric, or falling objects may follow entirely implausible trajectories (Motamed et al., 2026; Thozhiyoor et al., 2025; Kang et al., 2025). To address these hallucinations, recent efforts have attempted to force physical alignment by injecting simulators,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

physical priors, or physics-aware objectives into the training pipeline (Liu et al., 2024; Wang et al., 2025a; Pathak et al., 2026; Wang et al., 2026). However, these introduce a dilemma: they require costly post-training and consistently degrade the model's general generative capabilities, imposing a severe "alignment tax" for improving the physical correctness.

Must we train physics into these models, or is the knowledge already there? Recently, Gan & Isola (2026) showed that large language models (LLMs) contain "Neural Thickets", where dense neighborhoods of task-specific experts are located within the Gaussian search space of the pretrained weights. This raises a question for video models: Are physics experts already concealed within the weights of pretrained foundation video models?

The answer is far from obvious, given that the training paradigms of video models differ from those of LLMs. We discover that state-of-the-art video models contain specialized subnetworks that excel in different physical domains, such as Solid Mechanics, Fluid Dynamics, and Optics. We uncover these experts without post-training or weight updates, just by applying random binary masks to a small fraction of attention heads and feed-forward weights. Inspired by Gan & Isola (2026), we term these diverse populations of physics experts *physics thickets*. As shown in Figure 1, different masks reveal different combinations of physical strengths.

We compare these subnetworks with populations obtained by varying the base model's generation seed. In the largest model of each family, the highest-scoring subnetworks exceed the best of 100 sampled seeds in the mean Physics-IQ score. In *Wan2.2-A14B*, 78% of the randomly sampled subnetworks match or exceed the base model's 100-seed mean, and 48% improve on it by at least 5%, compared with 48% and 15% of the base model's seed draws. The same 100 Wan-A14B masks retain correlated overall and category-specific rankings in three generation seeds, showing reproducible differences between subnetworks across the tested seeds (Appendix B.2).

We also demonstrate the practical value of physics thickets by showing that high-performing subnetworks can be selected and reused without relying on ground-truth reference videos. By evaluating candidates using a video reward model on Physics-IQ or known constants (e.g., gravitational acceleration) on a free-fall benchmark, we can reliably isolate subnetworks that achieve higher mean physical accuracy and lower estimation errors than seed controls selected by the same screeners. On Physics-IQ, the initially selected masks retain their score advantage at two fresh generation seeds.

Finally, this paper reveals a surprising aspect in video models: obvious physical mistakes coexist alongside highly accurate, specialized physics experts. By exploring physics thickets, we offer a powerful, training-free way to unlock the physical knowledge already hidden within pretrained video generators.

Main Findings:

1. Random dropout uncovers diverse physics experts within pretrained video generators. By masking a small fraction of attention heads and feed-forward neurons, we find subnetworks good in different physical domains, without any training or weight update (Figure 2).

2. Solution density increases with model scale under the fixed masking rate. Larger models contain a greater fraction of subnetworks whose Physics-IQ scores match or exceed the seed-averaged base model, suggesting that scale makes high-performing physics subnetworks easier to find (Figure 3).

3. Physics subnetworks include both specialists and generalists. Many masks show domain-specific gains, indicating specialized behavior, while some masks improve performance across multiple physical domains, showing that generalist physics experts also emerge within the model (Figure 4).

4. Physics experts keep perceptual video quality while improving physical correctness. Human evaluation shows comparable visual fidelity and temporal consistency between selected subnetworks and generation-seed selection, with preference scores close to the equal-preference point of 50% (Table 3).

2 Sampling Subnetworks from Pretrained Video Generators

We search for physics experts by sampling structured perturbations of a pretrained generator. For each model, we sample $M = 100$ subnetworks by zeroing selected attention heads and feed-forward neurons. A binary *mask* m_i specifies which units are zeroed; a *subnetwork* is the resulting generator with the remaining pretrained weights unchanged, following the usual masked-network usage (Frankle & Carbin, 2019; Ramanujan et al., 2020). Each mask is fixed across scenes and applied only at the prescribed denoising stages. This construction is a form of structured dropout at inference time. The resulting population is

$$\mathcal{P}_{\text{sub}} = \{G_{\theta,m_i}(\cdot; z_0)\}_{i=1}^{M}, \tag{1}$$

where G_{θ,m_i} denotes the generator with mask m_i and the prescribed denoising-stage schedule. We evaluate the initial subnetwork population and its unmasked base on the same Physics-IQ tasks at an anchor generation seed z_0. For each of the five models, we also evaluate 100 unmasked generation-seed candidates and use their mean score as the solution-density reference. As a robustness check, we evaluate the same 100 Wan-A14B masks at the anchor seed and two fresh generation seeds (Appendix B.2). The main population comparisons use each model's anchor-seed results. Appendix B provides additional analyses of the subnetwork and generation-seed populations.

We study Wan2.2-TI2V-5B and Wan2.2-I2V-A14B (Team Wan et al., 2025; Team Wan, 2025), and Cosmos3-Edge, -Nano and -Super (NVIDIA, 2026). Both Wan models use an 8% masking rate for attention heads and FFN neurons during high-noise denoising: we mask the high-noise expert in A14B and apply the mask at timesteps $t \geq 900$ in 5B, targeting comparable denoising stages. All three Cosmos models use a 4% masking rate for attention heads and FFN neurons in both the understanding and generation pathways, applied at timesteps $t \geq 900$. Appendix A gives implementation details.

We evaluate physical behavior using Physics-IQ (Motamed et al., 2026), which compares generated videos with real reference videos. The benchmark covers 198 tasks across five physical categories, with three camera views for each of 66 scenarios. Scores measure motion overlap and pixel differences, accounting for variation between real recordings. We report each candidate's average score across tasks.

3 Large Generators Contain Dense Physics Thickets

We begin with the question: how frequently does random masking reveal a useful physics subnetwork? On Physics-IQ, we measure solution density across model scales and physical categories, then compare the resulting subnetwork populations with equally sized populations obtained by varying the base model's generation seed.

Figure 2 shows 100 sampled subnetworks for each Cosmos3 model through scores in the three largest physical categories and an RGB composite. Appendix B.3 gives the corresponding Wan maps and rendering details; Appendix B.1 gives solution density for all five categories.

3.1 Solution density across model scales and physical categories

For each model and task group T, let $Q_{i,T}$ be subnetwork i's mean Physics-IQ score and $\bar{B}_T$ the base model's mean over 100 generation seeds (Section 3.3). Averaging reduces sensitivity to the choice of seed (Appendix B.2). We estimate solution density as the fraction of sampled subnetworks scoring at least $a\bar{B}_T$:

$$\widehat{\delta}_T(a) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}[Q_{i,T} \geq a\bar{B}_T], \qquad M = 100. \tag{2}$$

At $a = 1$, this counts subnetworks that match or exceed the base model's expected score; $a = 1.05$ requires at least a 5% improvement. The seed population provides an empirical reference at each threshold.

Larger models contain more subnetworks that match or improve on the base model. Within Wan2.2, A14B has higher overall solution density than 5B at every threshold (Figure 3). At $a = 1$,

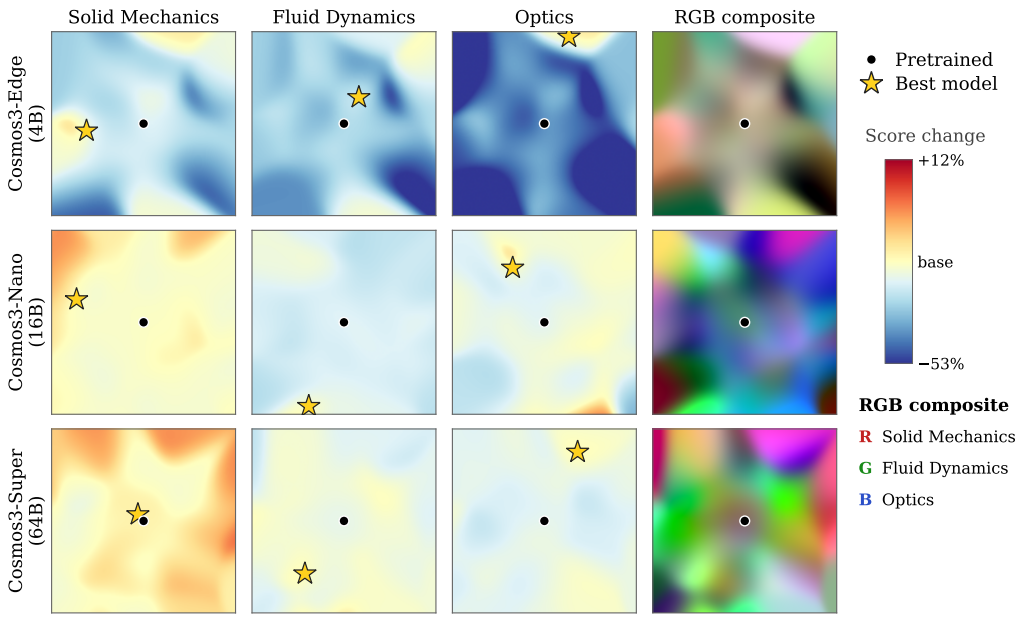

Figure 2: Physics thickets across Cosmos3 scales. Rows show Edge, Nano and Super; each row visualizes 100 subnetworks in a random projection of weight perturbations. These smoothed score maps are schematic; projection coordinates have no intrinsic geometric meaning. The three category maps show percentage score changes from the base model's mean score over its 100 generation seeds (Section 3.1) on a shared colour scale. Black dots mark the pretrained weights at the origin of the projection; large yellow stars mark the best evaluated subnetwork per category. RGB combines raw solid-mechanics (red), fluid-dynamics (green) and optics (blue) scores with independent channel scaling within each model. Wan maps and rendering details appear in Appendix B.3.

density rises from 1% to 78%, and 48% of A14B subnetworks improve on the base model's expected score by at least 5%, against 48% and 15% for the base model's own seed draws. Cosmos3 shows the same ordering: at $a = 1$, density rises from 7% in Edge to 27% in Nano and 58% in Super. The subnetwork population thus shifts from mostly damaging in the smaller models to mostly matching or improving in the largest model of each family. The largest gains on Wan-A14B occur in solid mechanics and optics; reproducible fluid experts are identified in Section 3.2. For Wan-A14B, evaluating the same 100 masks at two further generation seeds gives densities of 71% and 80% at $a = 1$ (Appendix B.2).

Greater robustness to dropout may contribute to the scale trend, but robustness alone predicts at most harmless masks, whose scores would match the seed population's. Wan-A14B exceeds this reference both at $a > 1$ and in its pool mean at all three tested seeds (Appendix B.2).

3.2 Single-task experts and generalists

We compare each subnetwork with the base model's mean score over 100 generation seeds in solid mechanics, fluid dynamics and optics. A *single-task expert* exceeds this baseline in exactly one of these categories; a *generalist* exceeds it in all three. Figure 4 shows examples of both. At the anchor generation seed, 19 of 100 Wan-A14B subnetworks are generalists; none of the 100 Cosmos-Super subnetworks meet this criterion. Appendix B.4 gives the selection rules and counts for all five models.

Category advantages transfer across scenarios and generation seeds. We identify category leaders using reference-based scores on one set of scenarios and evaluate them on held-out scenarios. Leaders selected for the target category outperform leaders selected for other categories by an average of 1.54 Physics-IQ points on Wan-A14B and 2.08 on Cosmos-Super (Appendix B.4). On

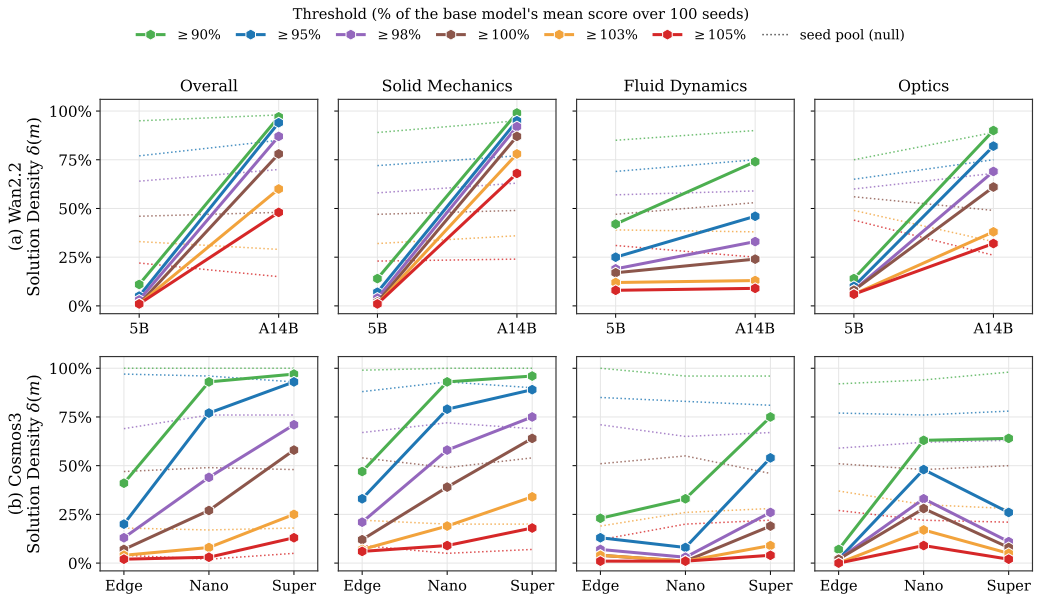

Figure 3: **Solution density across model scales and physical categories.** Rows: Wan2.2 and Cosmos3; columns: overall Physics-IQ, solid mechanics, fluid dynamics and optics. Solid curves give the fraction of 100 anchor-seed subnetworks scoring at least a times the base model's 100-seed mean; dotted curves give the corresponding seed-population fraction. Colors encode a as a percentage. Wan-A14B has higher density than Wan-5B across categories; overall density increases from Edge to Nano to Super. Lines connect evaluated models. Five-category results: Appendix B.1.

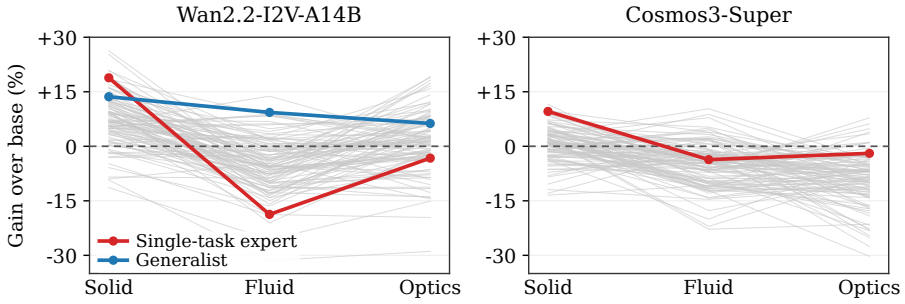

Figure 4: **Single-task experts and generalists.** Category scores relative to the base model's 100-seed average. Red: single-task expert; blue: generalist; gray: all 100 subnetworks at each model's anchor seed. The dashed line marks the baseline.

Wan-A14B, same-category ranks correlate 0.63–0.65 across generation seeds, compared with 0.10–0.13 across different categories. Nineteen masks reach the fluid baseline at all three tested seeds, and the generalist in Figure 4 remains above baseline in all three categories at both fresh seeds (Appendix B.2). Together, these results show reproducible category-dependent strengths in fixed subnetworks.

3.3 Eliciting expertise through masks and generation seeds

Generation-seed selection offers another route to eliciting physical expertise from the same pre-trained generator. We compare $\mathcal{P}_{\text{sub}}$ with $M = 100$ seed candidates evaluated on the same Physics-IQ tasks:

$$\mathcal{P}_{\text{seed}} = \{G_\theta(\cdot; z_i)\}_{i=1}^{M}. \tag{3}$$

Here, G_θ denotes the base generator and z_i is the generation seed for candidate i.

The two routes yield different candidate score distributions. Among the sampled candidates, ten of the 100 Wan-A14B subnetworks and five of the 100 Cosmos-Super subnetworks score above the maximum of the corresponding 100-seed population; the differences between sampled maxima are about 2.6 and 0.6 Physics-IQ points. The count depends on the generation seed: at two further seeds, 2 and 4 Wan-A14B subnetworks exceed the best seed, and two subnetworks do so at all three seeds (Appendix B.2). These observations characterize the populations reached under the specified sampling budgets and motivate exploring masks alongside generation seeds.

4 Finding and Reusing Physics Experts

We turn from population characterization to elicitation, comparing generation-seed selection and random-mask selection on Wan2.2-A14B and Cosmos3-Super. A common screening protocol tests the effect of the candidate population. We then examine whether the selected masks can be reused across scenes and generation seeds without reference videos.

4.1 Candidate Selection

Random-mask selection (*Ours*) and generation-seed selection (*Seed BoN*) use the same generation budget, top-K screening and final-video selection rule. This comparison tests the effect of the candidate population under a common screening procedure. Selected candidates are reused across all evaluation scenes, retaining one output per scene. We also compare with Proprio (Hassan et al., 2026), which selects candidates using a latent self-score, matching generator-inference FLOPs within 1% (Appendix A.3).

Physics-IQ. For each pool of $M = 100$ candidates, we use WMReward (Yuan et al., 2026) to rank candidates by their median V-JEPA-2 prediction surprise (Assran et al., 2025) over $d = 20$ tasks, with lower values preferred. We select one candidate ($K = 1$) for all 198 tasks. Selection requires no ground truth, and final scores cover all 198 tasks, reusing the selected candidate's videos from the 20 selection tasks.

Gravity. The gravity benchmark uses 50 two-ball free-fall scenes from Thozhiyoor et al. (2025). We use the benchmark's SAM 3 detector (Carion et al., 2025) to track each ball; its known release height h and measured flight time t give $\hat{g} = 2h/t^2$. We evaluate gravity accuracy using capped log-RMSE:

$$L = \sqrt{\frac{1}{100}\sum_{b=1}^{100} e_b^2}, \qquad e_b = \begin{cases} \min\{|\ln(\hat{g}_b/9.81)|, \ln 6\}, & \text{both balls of the scene tracked,} \\ \ln 6, & \text{otherwise.} \end{cases} \tag{4}$$

We rank $M = 200$ candidates by their mean capped absolute log-gravity error on 30 single-ball scenes, using the known release heights and ground-truth acceleration $g = 9.81\,\mathrm{m/s^2}$. We evaluate the selected candidates on 50 separate two-ball scenes. The top $K = 16$ candidates each generate a video for every evaluation scene; a vote based on their estimated accelerations selects the final video. Appendix D describes the evaluation protocol.

Physical behavior of selected subnetworks. Figure 5 illustrates physical strengths across solid mechanics, optics, magnetism and fluid dynamics, reusing one selected mask per model across scenes.

4.2 Selection Results and Human Evaluation

We run Ours independently at three generation seeds per model, reselecting subnetworks in each run. The Seed BoN result is shared across these runs. For Proprio, we follow the published scoring settings, evaluate global, motion and hybrid scoring, and report the best variant for each model and benchmark (Appendix A.3).

Physics-IQ. Ours scores 39.00 on Wan and 43.55 on Cosmos, exceeding Proprio's 37.57 and 42.21 under matched FLOPs (Table 1). Improvements over Seed BoN are $+2.75$ and $+2.08$ points, and persist on the 141 tasks from scenarios excluded from screening ($+2.75$ and $+1.69$; Appendix C).

(a) Solid mechanics
Wan2.2-A14B

Prompt: A 30lb kettlebell and a green piece of paper are lowered onto two pillows. Static shot with no camera movement.

(b) Optics
Wan2.2-A14B

Prompt: A yellow mug and a blue wooden block are placed on a rotating turntable that rotates clockwise illuminated by a spotlight casting their shadow on the wall behind it. Static shot with no camera movement.

(c) Magnetism
Cosmos3-Super

Prompt: A powerful magnet is placed on the table facing a black rotating platform that rotates clockwise. A small white plastic domino block is placed on the platform and is rotating towards the magnet. Static shot with no camera movement.

(d) Fluid dynamics
Cosmos3-Super

Prompt: A black balloon is attached to a fixed electric air pump hose on a table with a plain wall in the background. Air is being pumped in the balloon. Static shot with no camera movement.



Figure 5: **Physical behavior elicited across categories.** (a) Ours better captures the effect of weight: the heavy kettlebell compresses its pillow more than the light sheet of paper, whereas Seed BoN produces similar deformation under both objects. (b) Ours better preserves the distinct mug and narrow block shadows during rotation, whereas Seed BoN develops a widened, distorted block shadow. (c) Ours keeps the plastic domino moving with the turntable; Seed BoN draws it toward the magnet. (d) Ours shows progressive balloon expansion under air inflow, while Seed BoN remains nearly constant in size. Strips show four frames over 0–5 s. One fixed candidate per method and model is reused across scenes.

Table 1: Physics-IQ (↑) and human preference on 198 tasks, including 20 screening tasks. Ours: mean ± sample s.d. over three seeds; Seed BoN is shared. Δ = Ours − Seed BoN. Human Win excludes ties; Pref. counts ties as half a win (%, 594 task–seed pairs per model). Ours/Seed BoN use $M = 100$; Proprio uses $M = 83/86$ on Wan/Cosmos to match FLOPs. All use $K = 1$. Per-seed results and human protocol: Appendices C and F.

Method / measure	Wan2.2-A14B	Cosmos3-Super
Seed BoN	36.25	41.47
Proprio	37.57	42.21
Ours	**39.00** ±1.62	**43.55** ±1.21
Δ (↑)	+2.75	+2.08
Human preference: Ours vs. Seed BoN		
Win (%) ↑	56.0	56.3
Pref. (%) ↑	53.5	53.9

Table 2: Gravity capped log-RMSE (↓) and human preference on 50 two-ball scenes, with top-16 voting. Ours: mean ± sample s.d. over three seeds; Seed BoN is shared. Δ = Ours − Seed BoN before rounding. Human Win excludes ties; Pref. counts ties as half a win (%, 150 scene–seed pairs per model). Ours/Seed BoN use $M = 200$; Proprio uses $M = 168/174$ on Wan/Cosmos to match FLOPs. Per-seed results and human protocol: Appendices D.4 and F.

Method / measure	Wan2.2-A14B	Cosmos3-Super
Seed BoN	1.233	0.729
Proprio	1.224	**0.660**
Ours	**1.005** ±0.151	0.687 ±0.098
Δ (↓)	−0.229	−0.042
Human preference: Ours vs. Seed BoN		
Win (%) ↑	63.6	57.0
Pref. (%) ↑	61.7	56.0

Gravity. On gravity, Ours yields mean log-RMSEs of 1.005 on Wan and 0.687 on Cosmos, compared with 1.233 and 0.729 for Seed BoN (Table 2). Compared with Proprio (1.224 on Wan and 0.660 on Cosmos), Ours has lower gravity error on Wan and slightly higher error on Cosmos. Appendix D.4 reports the per-run results.

Physical plausibility evaluation. We conduct a pairwise human evaluation with 30 participants, comparing Ours and Seed BoN on the same input scene and prompt. On Physics-IQ, evaluators judge which video depicts more physically plausible behavior while following the prompt. On gravity, they judge which video depicts falling motion more consistent with real-world gravity. Human judgments favor Ours across both benchmarks and model families, with win rates of 56.0–63.6% (Tables 1 and 2). Appendix F gives the protocol and scoring definitions.

Visual quality evaluation. We also compare the visual fidelity and temporal consistency of outputs elicited by the two methods. For each benchmark and model, we randomly sample 30 outputs from each of the three mask re-selection runs and pair each with the corresponding Seed BoN output, yielding 90 matched video pairs. Evaluators judge the two criteria separately, choosing either video or no noticeable difference. Preference scores remain close to 50% (49.4–51.1%), indicating comparable visual quality between Ours and Seed BoN (Table 3). Appendix F.2 gives the protocol and win, loss and tie counts.

4.3 Reusing subnetworks at new generation seeds

We reuse the initially selected subnetwork on Physics-IQ and the initially selected top-16 set on gravity at two fresh seeds, without repeating selection. The transferred Physics-IQ subnetworks

Table 3: Visual quality evaluation. Ours versus the final Seed BoN outputs, with 90 matched pairs per benchmark and model (30 per generation run). Both columns report preference scores (%, ↑), counting ties as half a win; 50% denotes equal aggregate preference. Visual fidelity and temporal visual consistency are judged separately.

Benchmark	Model	Visual fidelity Pref. (%)	Temporal consistency Pref. (%)
Physics-IQ	Wan2.2-A14B	49.4	51.1
	Cosmos3-Super	50.0	51.1
Gravity	Wan2.2-A14B	50.6	50.0
	Cosmos3-Super	49.4	50.6

score 39.39–39.51 on Wan and 42.91–43.68 on Cosmos, compared with 36.25 and 41.47 for Seed BoN. The selected Physics-IQ masks therefore retain their advantage at both fresh seeds without additional screening. On gravity, the transferred masks have lower error than Seed BoN at both fresh seeds on Wan and at one on Cosmos. Tables 13 and 17 report all runs and re-selection results.

5 Related Work

Physics and inference-time selection. Simulator-driven generation (Liu et al., 2024; Zhang et al., 2024; Wang et al., 2025a), physical conditioning (Satish et al., 2026; Pathak et al., 2026; Zhang et al., 2025) and physics-aware objectives (Wang et al., 2026) address physical consistency through added supervision or structure. Video test-time search (Liu et al., 2025; Kim & Kim, 2026) and latent-world-model scoring (Yuan et al., 2026) select outputs; Proprio (Hassan et al., 2026) uses the generator's own latent residuals for self-scoring. We study random-mask selection alongside generation-seed selection as routes to eliciting physical expertise, with a focus on the prevalence, specialization and reuse of subnetworks already present in pretrained weights.

Subnetworks and component interventions. Lottery tickets and supermasks establish useful subnetworks in trained and random networks (Frankle & Carbin, 2019; Zhou et al., 2019; Ramanujan et al., 2020). Head pruning (Michel et al., 2019; Voita et al., 2019), motion-head analysis (Jung et al., 2026), and randomized component attribution (Ghorbani & Zou, 2020) motivate structured interventions. Neural Thickets (Gan & Isola, 2026) studies nearby experts through Gaussian weight perturbations in language models. We investigate the thicket phenomenon in video generators by sampling subnetworks, comparing them with equally sized sampling-seed populations. MaskUNet (Wang et al., 2025b) learns sample- and timestep-dependent masks for image diffusion; we select subnetworks whose masks are fixed across scenes. Guidance with weakened model copies (Hyung et al., 2025; Chen et al., 2026; Karras et al., 2024) runs the altered copy alongside the base model, whereas our subnetwork is itself the generator.

6 Discussion

Physics thickets reveal reusable physical expertise within pretrained video generators. The experiments connect population structure to practical elicitation: larger evaluated models contain denser populations of useful subnetworks, category advantages transfer to held-out scenarios, and mask rankings reproduce across the tested generation seeds. Screening these populations yields higher mean Physics-IQ scores and lower mean gravity errors than Seed BoN under matched budgets. Against Proprio at matched FLOPs, Ours scores higher on Physics-IQ for both models and has lower gravity error on Wan; Proprio has lower gravity error on Cosmos. Human judgments favor the selected subnetworks' physical plausibility while finding comparable visual quality.

The component analysis links Wan's gravity accuracy to which attention heads are removed: a head set selected by attribution achieves lower gravity error than random sets matched in size and layer distribution (Appendix E). Fixed masks provide an explicit, reusable way to access these strengths without updating the pretrained weights.

AI USE STATEMENT

We used generative AI tools to assist with drafting and polishing the manuscript, identifying and retrieving relevant literature, providing feedback on the interpretation of experimental results, and checking submission requirements. The authors are responsible for verifying all AI-assisted content and take full responsibility for the accuracy and integrity of the final manuscript, including its text, scientific claims, reported results, and references.

ETHICS STATEMENT

The human evaluation study described in Appendix F received an exemption from institutional review board (IRB) review. The institution is anonymized for double-blind review.

REPRODUCIBILITY STATEMENT

The appendices specify mask construction, screening, voting, scoring conventions and the reported comparisons. Candidate-level scores, configuration manifests and scoring code are intended for release.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models. In *International Conference on Learning Representations*, 2026.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Yulu Gan and Phillip Isola. Neural thickets: Diverse task experts are dense around pretrained weights, 2026. URL https://arxiv.org/abs/2603.12228.

Amirata Ghorbani and James Zou. Neuron Shapley: Discovering the responsible neurons. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2002.09815.

Mariam Hassan, Kaouther Messaoud, Wuyang Li, and Alexandre Alahi. Proprio: Latent self-scoring and inference-time refinement for physically plausible video generation, 2026.

Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Sunyoung Jung, Jiwoo Park, Yoonseok Choi, Kyobin Choo, Ming-Hsuan Yang, and Seong Jae Hwang. Controlling motion transfer in diffusion transformers via attention heads. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2607.11081.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Kwanyoung Kim and Sanghyun Kim. Model already knows the best noise: Bayesian active noise selection via attention in video diffusion model. In *International Conference on Learning Representations*, 2026.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-t1: Test-time scaling for video generation, 2025. URL https://arxiv.org/abs/2503.18942.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2409.18964.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI, 2026.

Saurabh Pathak, Elahe Arani, Mykola Pechenizkiy, and Bahram Zonooz. PhysVid: Physics aware local conditioning for generative video models. *arXiv preprint arXiv:2603.26285*, 2026.

Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11893–11902, 2020.

Siddarth Nilol Kundur Satish, Devesh Jaiswal, Hongyu Chen, and Abhishek Bakshi. PhysVideo-Generator: Towards physically aware video generation via latent physics guidance. *arXiv preprint arXiv:2601.03665*, 2026.

Team Wan. Wan2.2: Official model repository. GitHub repository, 2025. URL https://github.com/Wan-Video/Wan2.2.

Team Wan, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models, 2025.

Varun Varma Thozhiyoor, Shivam Tripathi, Venkatesh Babu Radhakrishnan, and Anand Bhattad. Objects in generated videos are slower than they appear: Models suffer sub-earth gravity and don't know Galileo's principle...for now, 2025.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. PhysCtrl: Generative physics for controllable and physics-grounded video generation. *arXiv preprint arXiv:2509.20358*, 2025a.

Cong Wang, Hanxin Zhu, Jiayi Luo, Yonglin Tian, Xiaoqian Cheng, Peiyan Tu, Xin Jin, Long Chen, and Zhibo Chen. Physics-informed video generation via Mixture-of-Experts latent alignment. *arXiv preprint arXiv:2606.04737*, 2026.

Lei Wang, Senmao Li, Fei Yang, Jianye Wang, Ziheng Zhang, Yuhan Liu, Yaxing Wang, and Jian Yang. Not all parameters matter: Masking diffusion models for enhancing generation ability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. URL https://arxiv.org/abs/2505.03097.

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16118–16129, 2026.

Haoze Zhang, Tianyu Huang, Zichen Wan, Xiaowei Jin, Hongzhi Zhang, Hui Li, and Wangmeng Zuo. PhysChoreo: Physics-controllable video generation with part-aware semantic grounding. *arXiv preprint arXiv:2511.20562*, 2025.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2404.13026.

Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1113d7a76ffceca1bb350bfe145467c6-Abstract.html.

A PROTOCOL

A.1 MASK SPECIFICATION

Each subnetwork is specified by a mask that zeroes whole units: an attention head or a feed-forward neuron. Wan head masks zero the corresponding rows of q, k, v and matching columns of o; Cosmos uses the grouped-query-safe convention specified below. Masks are sampled independently and held fixed across scenes and both classifier-free-guidance branches; the same mask is applied at every active denoising step of the specified stage. Nominal masking rates are applied per block, with masked counts rounded to whole attention heads and feed-forward neurons. In compact plot labels, "mask" denotes the corresponding subnetwork.

Wan comparisons. We align the generation task, denoising stage, masking rate and evaluation protocol to make the two released Wan checkpoints as comparable as possible. Both are evaluated in image-to-video mode on the same 198 Physics-IQ tasks, using the same initial images and prompts; the TI2V capability of 5B does not change the task evaluated here. A14B routes steps with $t \geq 900$ to its high-noise expert. We therefore mask that expert in A14B and apply the same timestep gate to the single-stream 5B model, leaving lower-noise denoising unmasked in both. This matches the stage of intervention despite the different module organization. Both models use the same structured removal of attention heads and FFN neurons at a nominal 8% per block. Finally, each checkpoint is evaluated with 100 masks and 100 unmasked generation seeds, and solution density uses the same relative thresholds against that checkpoint's 100-seed mean. This measures how often masking preserves or improves each model's own performance, accounting for differences in baseline score and avoiding dependence on a single baseline seed.

Cosmos3. For all three Cosmos3 models the masks act on both the understanding and generation pathways at timesteps $t \geq 900$, and are grouped-query-attention safe: they zero query rows of q and the matching columns of o, never k or v, and SwiGLU neurons. The Physics-IQ and gravity pools are draws from the same 4% family. The Cosmos3 rate is not matched to the Wan rate; scaling comparisons are made within each family.

Seed populations. Seed candidates use the unmasked checkpoint and differ only in the generation seed. The anchor seed of each subnetwork population is a member of the corresponding seed population, so every subnetwork can be compared with its own base model at its own seed; solution density is referred to the seed population's mean score (Section 3.1).

A.2 MATCHED COMPUTE BUDGETS

Within each model and benchmark, Ours and Seed BoN use the same candidate-pool size M, screening scenes, selected-candidate count K, and evaluation scenes, with the same generation settings and scoring procedure. Both therefore generate and score the same number of videos per comparison. Masking retains dense inference, and neither method requires additional training.

A.3 PROPRIO COMPARISON AND FLOPS ACCOUNTING

We evaluate the inference-time selection setting of Proprio (Hassan et al., 2026) on the same checkpoints and benchmarks, following its published scoring settings. The score uses four mid-noise levels spanning 0.3–0.6 of the sampling schedule, four noise perturbations per level, and inverse-variance weighting. We run all three scoring variants: S_{global}, S_{motion}, and S_{hybrid}, with the published hybrid weight $\lambda = 0.2$. The main tables report the best result among these variants for each model and benchmark (highest Physics-IQ score or lowest gravity error); S_{global} is best in all four cases (Table 5). Proprio replaces the screening score while retaining our screening/evaluation splits, $K = 1$ selection for Physics-IQ, and top-16 voting for gravity.

Compute accounting. We match approximate generator-inference FLOPs within each model, including Proprio's scoring overhead. Using the $2PL$ estimate for one forward pass, where P is the active parameter count and L is the number of processed tokens, one generation with T denoising steps costs $2T$ forward passes: one conditional and one unconditional pass per step for classifier-free guidance. Wan uses $T = 40$ (80 passes), and Cosmos uses $T = 50$ (100 passes). One Proprio score requires $4 \times 4 = 16$ forward passes, equivalent to $\rho = 0.20$ generations on Wan or 0.16 on

Cosmos. Our multiplicative masks zero units but retain dense execution, so we apply no FLOPs discount to masked generation. This accounting covers generator forward passes; external screening and evaluation models are outside this proxy.

Let M_0 be our candidate count, d the number of screening scenes, and E the subsequent evaluation generations. In units of one full generation, the budgets are

$$C_{\text{ours}} = M_0 d + E, \qquad C_{\text{Proprio}} = M d(1+\rho) + E. \tag{5}$$

For Physics-IQ, $M_0 = 100$, $d = 20$, and $E = 198 - 20 = 178$, since selection videos are reused. For gravity, $M_0 = 200$, $d = 30$, and $E = 16 \times 50 = 800$. We adjust Proprio's candidate count to approximate the corresponding budget. Table 4 records the counts used in the experiments; the resulting total budgets differ from ours by less than 1% in every case.

Table 4: Generator-inference budgets in full-generation equivalents, matched within each model. M is Proprio's candidate count; the final column is its total-budget difference relative to Ours.

Benchmark / model	Ours	Proprio	M	Difference
Physics-IQ / Wan	$100 \times 20 + 178$	$83 \times 20 \times 1.20 + 178$	83	-0.37%
Physics-IQ / Cosmos	$100 \times 20 + 178$	$86 \times 20 \times 1.16 + 178$	86	-0.22%
Gravity / Wan	$200 \times 30 + 800$	$168 \times 30 \times 1.20 + 800$	168	$+0.71\%$
Gravity / Cosmos	$200 \times 30 + 800$	$174 \times 30 \times 1.16 + 800$	174	$+0.81\%$

Table 5: All three Proprio scoring variants under the budgets in Table 4. Bold marks the best variant in each column, as reported in the main text.

	Physics-IQ ↑		Gravity log-RMSE ↓	
Proprio variant	Wan	Cosmos	Wan	Cosmos
S_{global}	**37.57**	**42.21**	**1.224**	**0.660**
S_{motion}	34.72	40.02	1.303	0.760
S_{hybrid}	37.26	40.02	1.295	0.760

A.4 Screeners

Both screeners below are computed without reference videos.

WMReward. We use the WMReward method introduced by Yuan et al. (2026), which scores physical plausibility using V-JEPA-2 prediction surprise. We use the official V-JEPA-2 ViT-g configuration at 256 px, window 16, context 8, stride 8, up to 120 frames, so that the whole clip is scored (81 frames for Wan, 120 for Cosmos). Candidates are ranked by median surprise over the screening scenes, ascending.

Gravity screener. On each of 30 single-ball scenes rendered in Blender outside the benchmark, the screener measures $\hat{g}$ from the tracked fall and averages $|\ln(\hat{g}/9.81)|$, capped at $\ln 6$. Clips in which SAM 3 detects more than one ball are dropped from the screening average; the exclusion applies to selection only and every benchmark scene is scored at evaluation. This screener reads the known release heights and the value of g as supervision; the Physics-IQ screener uses the video reward model. Both operate without reference videos.

B Physics-IQ populations

Aggregate scores use all 198 tasks. Category means use 114 solid-mechanics, 45 fluid-dynamics, 24 optics, 9 thermodynamics and 6 magnetism tasks.

B.1 Solution density

Figure 6 extends the density comparison to all five categories. On Wan-A14B, gains are most frequent in solid mechanics and optics. At the anchor seed, 24% of masks and 53% of seed draws

reach the fluid baseline; Appendix B.2 identifies the fluid experts whose advantage persists across generation seeds.

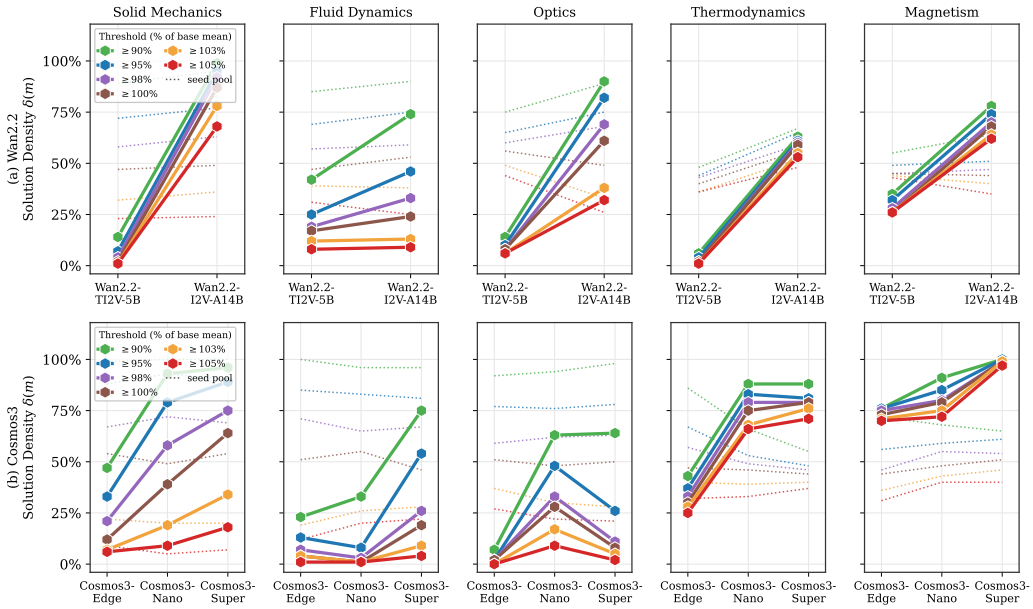

Figure 6: Solution density by physical category, in the format of Figure 3 and on all 198 tasks: within each panel the horizontal axis is model scale and each solid line corresponds to one value of a, with scores compared to the base model's mean score over its 100 generation seeds in that category; dotted lines give the same fraction for the seed draws themselves. (a) Wan2.2, (b) Cosmos3. Thermodynamics and magnetism contain 9 and 6 tasks respectively and are correspondingly less stable.

B.2 Robustness to generation seeds

We test whether Wan-A14B's population-level advantage and mask rankings reproduce across generation seeds. The 100 Wan-A14B masks were regenerated and evaluated on all 198 tasks at the Fresh A and Fresh B generation seeds in Table 13, in addition to the anchor seed. The seed population is the same 100 seeds throughout.

Reference for solution density. Table 6 shows why density is referred to the seed population's mean rather than to the base model's score at the anchor seed. The base model's single-seed scores in the anchor, Fresh A and Fresh B runs lie at the 87th, 90th and 31st percentiles of the seed population, and the fraction of masks matching or exceeding that single draw is 38%, 17% and 88% respectively. Referred to the population mean (35.19), the same three pools give 78%, 71% and 80% at $a = 1$ and 48%, 25% and 34% at $a = 1.05$; the seed draws themselves give 48% and 15%. Averaging each mask's score over the three seeds gives 82% and 41%. The subnetwork pool mean exceeds the seed pool mean at every seed ($+1.44$, $+0.74$ and $+1.29$; one-sided permutation $p < 10^{-3}$ in each case).

Mask quality is reproducible across generation seeds. The Spearman correlation between the 100 per-mask scores at two seeds is 0.70 (anchor–Fresh A), 0.69 (anchor–Fresh B) and 0.67 (Fresh A–Fresh B) (Figure 7). Two masks exceed the best of the 100 unmasked seeds at all three tested seeds. One of these masks is also selected by WMReward in the initial run (Section 4.2), linking population-level reproducibility to reuse of a selected subnetwork. These results support stable differences between masks across the tested seeds; they do not require each mask's score to be invariant to the generation seed.

Category strengths reproduce across generation seeds. Table 7 follows the same masks across three generation seeds. Within-category rank correlations are 0.63–0.65, compared with 0.10–0.13

Table 6: Wan-A14B: the same 100 masks at three generation seeds, all 198 tasks. The 100-seed population has mean 35.19 and maximum 39.26. $\widehat{\delta}$ uses the seed-population mean as reference (Equation 2); $\widehat{\delta}_{\text{own}}(1)$ uses the base score at that seed instead. Δ is the subnetwork-minus-seed mean difference on the evaluated tasks. The last row averages each mask's score over the three seeds.

Run	Mean $\pm$ s.d.	Max	$\widehat{\delta}(1.00)$	$\widehat{\delta}(1.05)$	$\widehat{\delta}_{\text{own}}(1)$	>best seed	Δ
Anchor	36.63 ± 2.16	41.82	78%	48%	38%	10	$+1.44$
Fresh A	35.93 ± 1.64	39.53	71%	25%	17%	2	$+0.74$
Fresh B	36.48 ± 1.68	41.42	80%	34%	88%	4	$+1.29$
Three-seed mean	36.35 ± 1.67	40.74	82%	41%	—	—	$+1.16$

across different categories. Nineteen masks reach the fluid-dynamics seed-population mean at all three tested seeds, identifying fluid experts whose above-baseline performance reproduces across these seeds.

Table 7: Wan-A14B category strengths across three generation seeds, using the same 100 masks. Spearman correlations compare within-population category ranks across seeds and are averaged over seed pairs. The last column counts masks at or above the unmasked 100-seed category mean at every tested seed.

	Cross-seed rank correlation		Above baseline
Category	Same category	Other category	at all three seeds
Solid mechanics	0.64	0.10	61 / 100
Fluid dynamics	0.63	0.11	19 / 100
Optics	0.65	0.13	31 / 100

Together, these results demonstrate that Wan-A14B's population-level advantage and category-specific mask rankings reproduce across all three tested generation seeds.

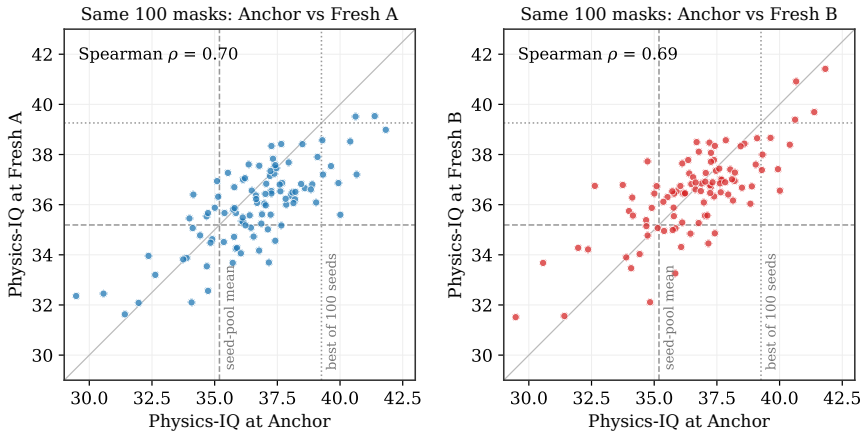

Figure 7: Wan-A14B: mean Physics-IQ of the same 100 masks in the anchor run against Fresh A (left) and Fresh B (right). Dashed lines mark the seed population's mean, dotted lines its maximum.

B.3 Wan landscapes and projection details

Projection and smoothing. Within each model, all panels share a fixed two-dimensional random projection of weight perturbations $\Delta\theta = -\theta \odot m$, with the pretrained model at the origin. The projection is sampled independently of evaluation scores. Gaussian-kernel averages provide

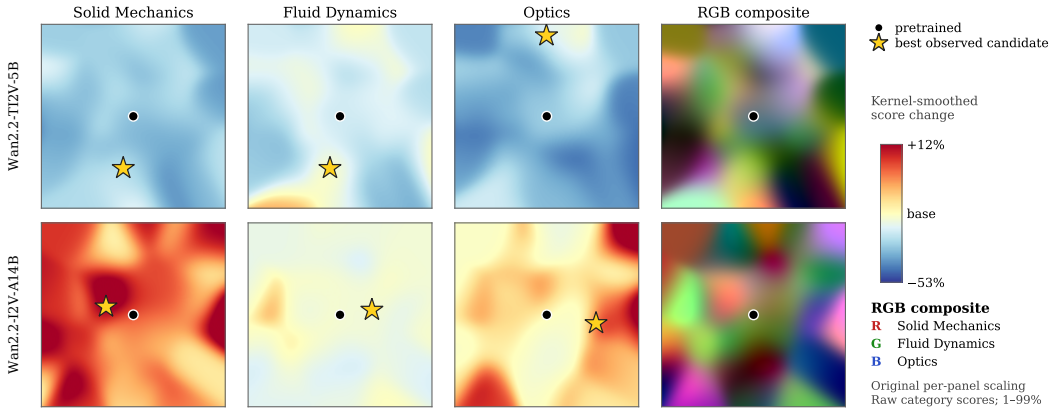

Figure 8: Physics thickets in Wan2.2, using the same rendering as Figure 2. These smoothed score maps are schematic; Rows show TI2V-5B and I2V-A14B. Category maps use the same score-change scale as Figure 2, with score changes measured from the base model's mean score over its 100 generation seeds. Black dots mark the pretrained weights at the origin of the projection; large yellow stars identify the highest-scoring evaluated subnetwork in each category. RGB channels are independently scaled within each model.

a smooth visual summary of the 100 evaluated subnetworks. A star identifies the highest-scoring observed subnetwork before smoothing; averaging neighbouring candidates can therefore move the maximum of the displayed field away from that star. Each model has its own projection coordinates.

RGB normalization. The red, green and blue channels are the kernel-smoothed raw scores for solid mechanics, fluid dynamics and optics, respectively. Each channel is linearly scaled between its own first and 99th percentiles within that model's smoothed grid and clipped to $[0, 1]$. Hue and brightness summarize relative category profiles within each model. Solution density is computed separately by counting evaluated subnetworks that meet the score threshold.

B.4 Cross-category improvements and rank reproducibility

Single-task experts and generalists. For category T, let $\bar{B}_T$ be the base model's mean Physics-IQ score over its 100 generation seeds and define the percentage gain of subnetwork i as

$$g_{i,T} = 100 \left(\frac{Q_{i,T}}{\bar{B}_T} - 1 \right). \tag{6}$$

A single-task expert has $g_{i,T} > 0$ in exactly one of the three categories; a generalist has $g_{i,T} > 0$ in all three. Scores are averaged over 114 solid-mechanics, 45 fluid-dynamics and 24 optics tasks. Counts use unrounded scores at each model's anchor seed.

Table 8: Generalists among 100 subnetworks per model. Baselines are category means over 100 unmasked generation seeds.

	Average baseline			Generalists
Model	Solid	Fluid	Optics	/ 100
Wan2.2-TI2V-5B	31.86	19.16	28.42	0
Wan2.2-I2V-A14B	36.46	26.33	48.69	19
Cosmos3-Edge	38.84	24.12	31.29	0
Cosmos3-Nano	44.65	24.81	40.01	0
Cosmos3-Super	47.09	25.75	50.31	0

In Figure 4, the red single-task experts exceed the baseline only in solid mechanics. Among these masks, the Wan-A14B example maximizes the solid-minus-fluid gain, and we show one solid-mechanics specialist for Cosmos3-Super. The blue generalist maximizes the minimum gain across

all three categories. On Wan-A14B, its gains are $+13.64\%$, $+9.32\%$ and $+6.28\%$ in solid mechanics, fluid dynamics and optics, respectively. This generalist remains above the seed-averaged baseline in all three categories at both fresh generation seeds, with gains ranging from $+3.2\%$ to $+7.5\%$ (Appendix B.2). No generalist is highlighted for Cosmos-Super because none meets the criterion.

Category ranks. For the rank-reproducibility analyses, we use $r_{i,T} = 100(k_{i,T} - 1)/99$, where $k_{i,T}$ is subnetwork i's ascending rank among the 100 candidates on category T, averaging tied ranks. These ranks measure within-population standing and are not used to define generalists.

Reproducibility across scenario subsets. For each of the three categories, we split scenarios into disjoint halves, keeping all views of a scenario in the same half, and rank subnetworks by mean score on each half. We summarize rank correlations between halves of the same category as r_{within} and between different categories as r_{across}. The analysis uses 183 tasks from 61 scenarios; thermodynamics and magnetism are excluded because of their small sample sizes. Point estimates average 2,000 random scenario splits; intervals come from 500 bootstrap resamples of the 100 subnetworks, each averaged over 200 splits. Table 9 reports the resulting correlation summaries, conditional on the evaluated tasks and anchor seed.

Table 9: Category standing across disjoint scenario subsets. r_{within} and r_{across} are rank correlations across 100 subnetworks between halves of the same category and of different categories (solid mechanics, fluid dynamics, optics). Corrected: cross-category correlation divided by the geometric mean of the corresponding split-half reliabilities, then averaged over category pairs. D: the geometric-mean reliability minus r_{across}, with 95% bootstrap intervals over subnetworks. Viewpoints: the corrected value when the three camera views of each scenario replace the categories.

Model	r_{within}	r_{across}	Corrected	D [95% CI]	Viewpoints
Wan2.2-TI2V-5B	0.44	0.31	0.71	0.121 [0.039, 0.189]	1.07
Wan2.2-I2V-A14B	0.32	0.12	0.42	0.176 [0.054, 0.269]	0.96
Cosmos3-Edge	0.70	0.61	0.86	0.097 [0.047, 0.142]	0.99
Cosmos3-Nano	0.31	0.14	0.45	0.157 [0.041, 0.246]	0.96
Cosmos3-Super	0.33	0.08	0.22	0.243 [0.125, 0.345]	1.08

Reliability correction. Unequal measurement reliability can lower cross-category correlations even when candidates share a common quality component. For each category pair A, B, we therefore compute

$$c_{AB} = \frac{r_{AB}}{\sqrt{r_A r_B}}, \qquad D_{AB} = \sqrt{r_A r_B} - r_{AB},$$

where r_A, r_B are the within-category split-half rank correlations and r_{AB} is the cross-category correlation. Correlations are averaged over splits before forming each ratio; the table then averages the three category-pair results. This attenuation-style correction is a diagnostic motivated by a common-factor model with independent measurement errors, under which c_{AB} is expected to be near 1.

Corrected category correlations are 0.42 for Wan-A14B and 0.22 for Cosmos-Super, with positive D intervals in all five models. Applying the same procedure to viewpoint groupings gives corrected correlations of 0.96–1.08 and $|D| \leq 0.04$. Category separation thus persists after reliability correction, while viewpoint groupings yield near-unity correlations, supporting specialization by physical category.

Held-out category-leader transfer. To test whether category strengths extend to scenarios excluded from selection, we split each category's scenarios into two halves, keeping all three views together: 19/19 scenarios for solid mechanics, 7/8 for fluid dynamics and 4/4 for optics. Within each candidate pool, the candidate with the highest reference-based score on one half is that category's leader. We evaluate it on the held-out half of every category.

Gains are measured relative to the corresponding candidate pool's mean on the same held-out tasks. Own averages the three leaders' gains in their selection category; Other averages the gains of leaders selected on the other two categories. Point estimates average 1,000 random splits. Intervals use

500 bootstrap resamples of the 100 candidates, each averaged over 200 splits, conditional on the evaluated scenarios.

Table 10 shows positive average category advantages on held-out scenarios: Own–Other gains are 1.54 and 2.08 points, with 95% intervals above zero in both models. On Wan-A14B, selecting on the anchor run and evaluating held-out scenarios at Fresh A and Fresh B gives own-category gains of 2.8 and 2.4 points and Own–Other differences of 1.5 and 1.4 points, respectively. Selecting the top five candidates per category preserves the ordering (subnetwork differences 1.83 and 1.76 points for Wan and Cosmos).

Table 10: Held-out category-leader transfer on Physics-IQ. Gains are points above the corresponding candidate pool's mean on the same held-out scenarios, averaged over three categories and 1,000 splits. Own: leader selected on that category. Other: leaders selected on the other two categories. Selection and testing use the anchor generation seed. Brackets are 95% candidate-bootstrap intervals conditional on the evaluated scenarios.

Model	Candidate pool	Own (↑)	Other	Difference (↑)
Wan-A14B	Subnetworks	$+2.73$ $[+1.74, +3.82]$	$+1.19$ $[+0.10, +1.89]$	$+1.54$ $[+0.75, +2.60]$
	Generation seeds	$+0.72$ $[-0.38, +1.61]$	-0.16 $[-0.78, +0.57]$	$+0.88$ $[-0.45, +2.08]$
Cosmos-Super	Subnetworks	$+2.61$ $[+1.70, +3.45]$	$+0.53$ $[-0.01, +1.43]$	$+2.08$ $[+0.85, +2.89]$
	Generation seeds	$+0.22$ $[-0.47, +0.93]$	$+0.28$ $[-0.11, +0.58]$	-0.06 $[-0.67, +0.74]$

C PHYSICS-IQ SCREENING

C.1 THE SCREENING GRID

Table 11 varies pool size M and screening-task count d under WMReward with $K = 1$, at the initial generation seed. For $M = 50$, scores are mean $\pm$ s.d. over 20 random candidate subsamples; $P(\text{mask} > \text{seed})$ is the fraction in which Ours exceeds Seed BoN. The $M = 100$ rows use the full pools. Differences are computed before rounding.

Table 11: Physics-IQ screening configurations, evaluated on all 198 tasks. $P(\text{mask} > \text{seed})$ is the fraction of candidate-pool subsamples for which Ours has the higher score.

Model	d	M	Ours	Seed BoN	Δ	$P(\text{mask} > \text{seed})$
Wan-A14B	10	50	39.93 ± 1.00	36.41 ± 0.71	$+3.53$	100%
	10	100	39.94	36.25	$+3.69$	—
	20	50	39.82 ± 1.09	36.62 ± 0.46	$+3.20$	100%
	20	100	40.62	36.25	$+4.37$	—
Cosmos-Super	10	50	44.33 ± 0.77	42.61 ± 0.65	$+1.72$	85%
	10	100	44.93	42.58	$+2.36$	—
	20	50	44.36 ± 0.86	41.02 ± 0.90	$+3.35$	100%
	20	100	44.93	41.47	$+3.46$	—

C.2 EVALUATION ON TASKS EXCLUDED FROM SELECTION

The scores above cover all 198 tasks, including the d used for selection, whose selected-candidate videos are reused for evaluation. Candidates are selected using WMReward. Table 12 repeats the comparison on 141 tasks, excluding all views of every scenario used for selection.

The scenario-disjoint evaluation assesses performance on scenarios excluded from selection; the candidate-population comparisons in Appendix B condition on the evaluated task set.

Table 12: Held-out evaluation. "Ours" re-runs selection at each of three generation seeds and reports mean ± sample s.d.; Seed BoN is the same screener applied to the seed pool. The scenario-disjoint split removes all three views of every scenario used for selection.

Model	Method	all 198	141 tasks (scenario-disjoint)
Wan2.2-A14B	Seed BoN	36.25	37.95
	Ours	39.00 ± 1.62	40.70 ± 2.02
	Δ	$+2.75$	$+2.75$
Cosmos3-Super	Seed BoN	41.47	40.50
	Ours	43.55 ± 1.21	42.19 ± 0.69
	Δ	$+2.08$	$+1.69$

C.3 Variation across independent runs

Table 13: Physics-IQ across three independent generation runs, all 198 tasks. "Transfer" reuses the initially selected subnetwork; "re-selection" screens the pool again in each run. The initial run uses the same candidate in both columns; Fresh A and Fresh B use new generation seeds. Seed BoN is shared across runs.

Model	Run	Seed BoN	Transfer	Re-selection
Wan2.2-A14B	Initial	36.25	40.62	40.62
	Fresh A		39.51	38.99
	Fresh B		39.39	37.38
Cosmos3-Super	Initial	41.47	44.93	44.93
	Fresh A		42.91	42.72
	Fresh B		43.68	43.00

D Gravity

D.1 Scoring and voting

We evaluate 50 two-ball scenes from Thozhiyoor et al. (2025), with known release heights and 33 frames on Wan2.2-A14B or 32 on Cosmos3-Super, at 16 fps. The released benchmark detector assigns landing to the frame before a tracked ball first stops descending; flight time and release height give $\hat{g} = 2h/t^2$. We report the root mean square of $\min\{|\ln(\hat{g}/9.81)|, \ln 6\}$ over 100 balls.

Both methods use the same consensus vote to select one representative video per scene from their top 16 candidates, using only the candidates' estimated accelerations.

D.2 Screen-to-test transfer

Table 14 compares screening and test errors across candidates and summarizes the selected sets before voting.

Table 14: Gravity, $M = 200$ per pool. ρ is the Spearman correlation between the screening error on the 30 single-ball scenes and the test error on the 50 two-ball scenes. "Selected" is the mean test error of the 16 top-screened candidates before voting.

Model	Pool	ρ (screen, test)	pool mean	selected mean
Wan2.2-A14B	Subnetwork	.666	1.308	1.123
	Seed	.047	1.378	1.402
Cosmos3-Super	Subnetwork	.511	1.313	1.059
	Seed	$-.027$	1.263	1.311

D.3 SWEEPS OVER CANDIDATE COUNT AND POOL DEPTH

We vary the deployment count K at fixed $M = 200$, and pool size M at fixed $K = 16$, using matched budgets for the two arms.

Table 15: Gravity K sweep, $M = 200$ per pool. Values are capped log-RMSE; Δ is Ours minus Seed BoN.

Model	K	Subnetwork	Seed	Δ
Wan2.2-A14B	1	1.056	1.159	$-.103$
	4	.971	1.173	$-.201$
	8	.938	1.236	$-.298$
	16	.924	1.233	$-.309$
	32	.926	1.254	$-.328$
	200	1.084	1.279	$-.195$
Cosmos3-Super	1	1.255	1.369	$-.114$
	4	.626	.865	$-.238$
	8	.553	.780	$-.227$
	16	.592	.729	$-.137$
	32	.583	.669	$-.086$
	200	.662	.618	$+.044$

Table 16: Pool-depth sweep at $K = 16$. Both pools use the same M; values are capped log-RMSE.

Model	Pool	M=16	50	100	200
Wan2.2-A14B	Subnetwork	1.043	.973	.922	.924
	Seed	1.186	1.217	1.193	1.233
Cosmos3-Super	Subnetwork	.791	.697	.633	.592
	Seed	.761	.703	.743	.729

These sweeps characterize sensitivity to K and M at the initial seed. The Cosmos difference reverses sign at $K = 200$.

D.4 VARIATION ACROSS INDEPENDENT RUNS

Table 17: Gravity log-RMSE across three independent generation runs, $K = 16$. "Re-selection" screens the 200-subnetwork pool again in each run; "transfer" reuses the initially selected top-16 set. Initial, Fresh A and Fresh B correspond to the runs in Table 13. Seed BoN applies the same procedure to a fixed 200-seed pool; its selected candidates and result are shared across the generation runs for the subnetworks.

Model	Run	Seed BoN	Re-selection	Transfer
Wan2.2-A14B	Initial	1.233	0.924	0.924
	Fresh A		1.178	1.225
	Fresh B		0.911	0.938
	mean	1.233	1.005 ± 0.151	1.029 ± 0.170
Cosmos3-Super	Initial	0.729	0.592	0.592
	Fresh A		0.787	0.760
	Fresh B		0.681	0.655
	mean	0.729	0.687 ± 0.098	0.669 ± 0.085

Re-selection lowers Wan error at all three seeds; the Cosmos difference changes sign. The three observed runs show variation in the magnitude and, for Cosmos, direction of the differences.

E COMPONENT-LEVEL ANALYSIS

We test Wan-A14B head interventions on the 50 two-ball scenes, using the same landing detector and generation seed for all arms.

E.1 ATTRIBUTION AND MATCHED HEAD-SET INTERVENTIONS

We rank 1,600 heads from 500 random masks by the marginal effect of dropping each head on mean error over 30 screening scenes. A half-mask fit predicts held-out mask errors with $r = 0.165$ (permuted-score null 0.004 ± 0.050, $p \leq 0.010$). We then generate videos with 16 predicted-harmful heads removed, 16 predicted-protective heads removed, or one of 20 random sets matched to the attributed set's per-block head counts (Table 18).

Table 18: Matched 16-head interventions in the first run. Errors use all 50 scenes, including failure penalties. The random row reports mean $\pm$ sample s.d. across 20 sets. Tracked counts are out of 100 balls.

Intervention	log-RMSE	Δ	Tracked
Unmasked model	1.291	—	88
16 attributed heads	1.020	-0.271	96
16 random heads, 20 matched sets	1.320 ± 0.089	$+0.029$	85.0
16 reverse-attributed heads	1.562	$+0.270$	69

The attributed set outperforms all 20 matched controls (range 1.159–1.452), supporting an effect of head identity in this setting.

F HUMAN EVALUATION

Every questionnaire in both studies below included golden questions with known answers to ensure annotation quality.

F.1 PHYSICAL PLAUSIBILITY

We compare the selected outputs of Ours and Seed BoN (Section 4) in a pairwise study with 30 evaluators. For each model, the Physics-IQ study covers all 198 tasks at each of three generation seeds (594 comparisons), and the gravity study covers the same 50 two-ball test scenes at each seed (150 comparisons). Evaluators view the two videos generated for the same input scene and prompt. For Physics-IQ, they judge which video depicts more physically plausible behavior while following the prompt; for gravity, which video's falling motion is more consistent with real-world gravity. For each pair, they either prefer one video or mark no preference.

Let W, L and T denote the numbers of wins, losses and ties for Ours. We report the win rate, which excludes ties, and the preference score, which counts a tie as half a win:

$$\text{Win} = 100\frac{W}{W+L}, \qquad \text{Pref.} = 100\frac{W+\frac{1}{2}T}{W+L+T}. \tag{7}$$

A preference score of 50% denotes equal preference. Table 19 reports counts and scores pooled across seeds, together with the per-seed preference scores.

F.2 VISUAL QUALITY

We randomly sampled 30 matched output pairs from each of the three re-selection runs, without excluding poor outputs, giving 90 pairs per benchmark and model (360 in total); for gravity, we used the final top-16-voted videos. The 30 evaluators assessed visual fidelity and temporal visual consistency separately. Method identities were hidden, and left/right placement and pair order were randomized. For each criterion, evaluators chose either video or no noticeable difference. Table 20 reports the resulting wins, losses and ties for Ours and the preference score defined in Eq. (7); nearly

Table 19: Human evaluation of physical plausibility, Ours versus Seed BoN. W, L and T are pooled over three generation seeds (594 comparisons for Physics-IQ and 150 for gravity per model). Win excludes ties; Pref. counts ties as half a win. The last column lists Pref. for the Initial / Fresh A / Fresh B runs.

Benchmark	Model	W	L	T	Win (%)	Pref. (%)	Per-seed Pref. (%)
Physics-IQ	Wan2.2-A14B	192	151	251	56.0	53.5	53.5 / 52.5 / 54.3
	Cosmos3-Super	205	159	230	56.3	53.9	55.3 / 52.8 / 53.5
Gravity	Wan2.2-A14B	82	47	21	63.6	61.7	58.0 / 56.0 / 71.0
	Cosmos3-Super	73	55	22	57.0	56.0	60.0 / 51.0 / 57.0

all pairs are ties under both criteria. These results indicate that our mask-based method maintains visual quality comparable to the generation-seed baseline (Seed BoN) in both visual fidelity and temporal consistency.

Table 20: Human evaluation of visual quality, Ours versus Seed BoN, over 90 matched pairs per benchmark and model (30 per generation run). W, L and T are wins, losses and ties for Ours; Pref. counts ties as half a win.

		Visual fidelity				Temporal consistency			
Benchmark	Model	W	L	T	Pref. (%)	W	L	T	Pref. (%)
Physics-IQ	Wan2.2-A14B	3	4	83	49.4	7	5	78	51.1
	Cosmos3-Super	3	3	84	50.0	4	2	84	51.1
Gravity	Wan2.2-A14B	1	0	89	50.6	4	4	82	50.0
	Cosmos3-Super	0	1	89	49.4	3	2	85	50.6