Paris 2.0: A Decentralized Diffusion Model for Video Generation

Ali Rouzbayani, Bidhan Roy, Marcos Villagra and Zhiying Jiang
Bagel Labs (bagel.com) 🤗bageldotcom/paris2

We present Paris 2.0, the first video generation model pre-trained through decentralized computation. Its training recipe builds upon Paris 1.0 [2], the first ever open-weight Decentralized Diffusion Model (DDM), which showed that image generation can be trained without a monolithic GPU cluster. However, temporally coherent video generation had remained an open problem under decentralized training, and Paris 2.0 closes it.
In low-resolution text-to-video training, against a monolithic model trained on the same data under a matched total compute budget, Paris 2.0 cuts Fréchet Video Distance (FVD) from 561.04 **to** 279.01**, a** $\sim 2.0\times$ **improvement, and lifts CLIP text-video similarity and aesthetic score.**

Keywords: Decentralized diffusion models, video diffusion, world models, expert routing, flow matching

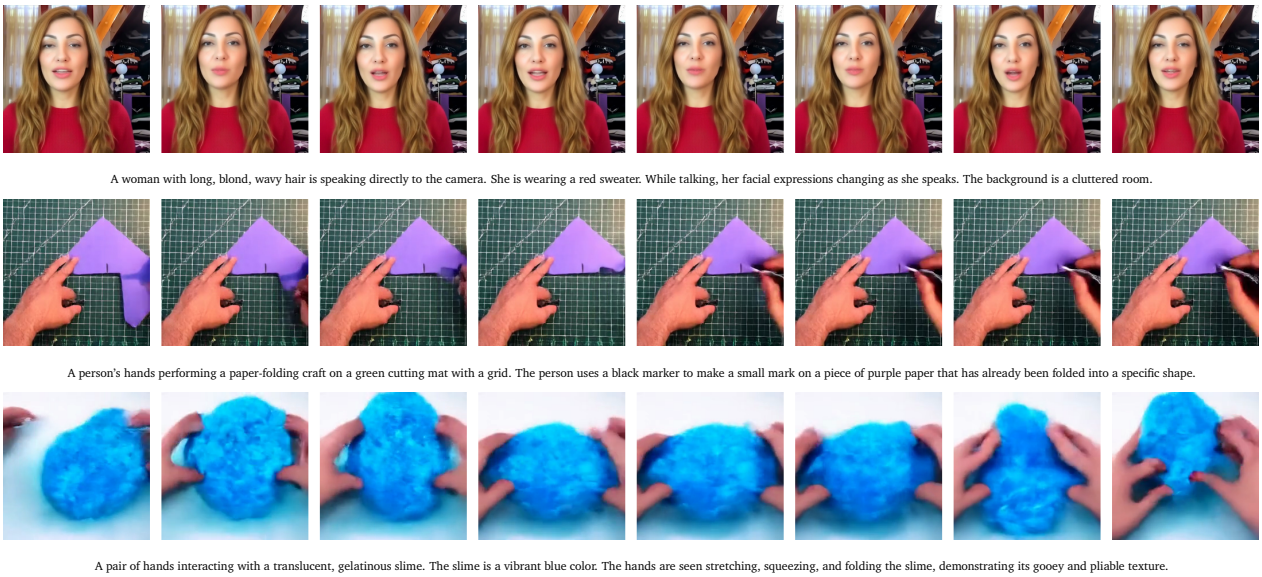

Figure 1 | Qualitative samples from Paris 2.0. Each row shows eight frames from one generated video.

1. Introduction

Video diffusion models are central to media generation, and the same backbones increasingly anchor the world models that physical AI agents roll forward to predict how their actions reshape a scene. The prevailing training paradigm nonetheless remains monolithic, in which a single model is trained on a broad mixture of video datasets and must subsequently generalize across all prompts at inference. Because the model's parameters are updated at the same time, the training infrastructure is bound to a homogeneous cluster of supply-constrained GPUs co-located behind high-bandwidth interconnects.

A Decentralized Diffusion Model (DDM) [6] removes that constraint. Paris 1.0 [2] demonstrated the approach for image generation, training an ensemble of diffusion experts without synchronization among them and routing across them at inference. However, whether the same recipe could yield

Corresponding email: research@bagel.com

temporally coherent video had remained an open question, since video burdens every expert with motion, longer temporal context, and substantially heavier latents. Paris 2.0 provides the affirmative answer. We extend the decoupled-compute training recipe to video and evaluate it directly against the monolithic alternative it aims to replace.

Under identical data, matched total training compute, and the same generation settings, a Stage 1 three-expert DDM surpasses the monolithic model on major video model benchmarks, as Figure 2 shows. Expert capacity trained in isolation does not merely survive the move to video, it outperforms the single backbone trained on the same data.

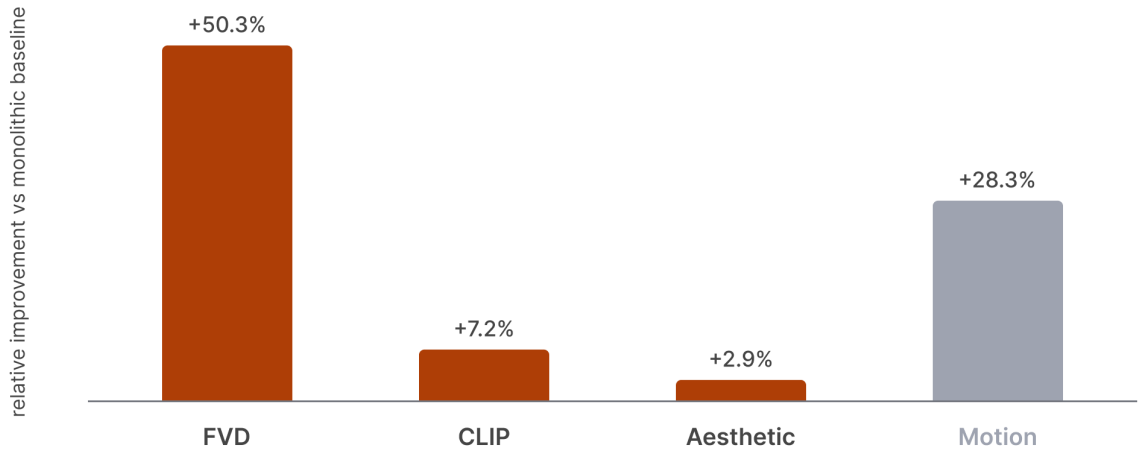

Figure 2 | Relative improvement over the monolithic baseline, where a taller bar marks a larger gain. Paris 2.0 roughly halves FVD and lifts CLIP text-video and aesthetic scores. Motion is a descriptive measure of per-frame displacement magnitude, not of temporal consistency, has no preferred direction, and is shown in gray. Absolute values appear in Table 1.

2. Decentralized Diffusion Models

A Decentralized Diffusion Model is an ensemble of independent diffusion experts, each a complete model trained on its own cluster of the data and combined at inference by a learned router. The experts exchange no gradients, parameters, or activations during training, so each is optimized in isolation, whereas monolithic training synchronizes across a tightly coupled pool of GPUs at every step. Eliminating this synchronization lifts the requirement that all compute be co-located, allowing each expert to train asynchronously on the least costly hardware available, including spot instances and eligible consumer hardware distributed across clouds and regions, as Figure 3 illustrates.

During inference, routing is performed at each denoising step, where a lightweight router reads the current noisy state and selects one or more experts to evaluate the video velocity field. Per-sample compute stays on the order of a single backbone while total capacity grows with each added expert, so the training problem becomes horizontally scalable, capacity is added by training another expert rather than by enlarging one synchronized run. A central and, to our knowledge, novel insight of Paris 2.0 is that this decomposition is especially well suited to video, where distinct clusters exhibit different motion patterns, camera behavior, and scene dynamics, so the router can exploit such specialization without making every sample pay for every expert.

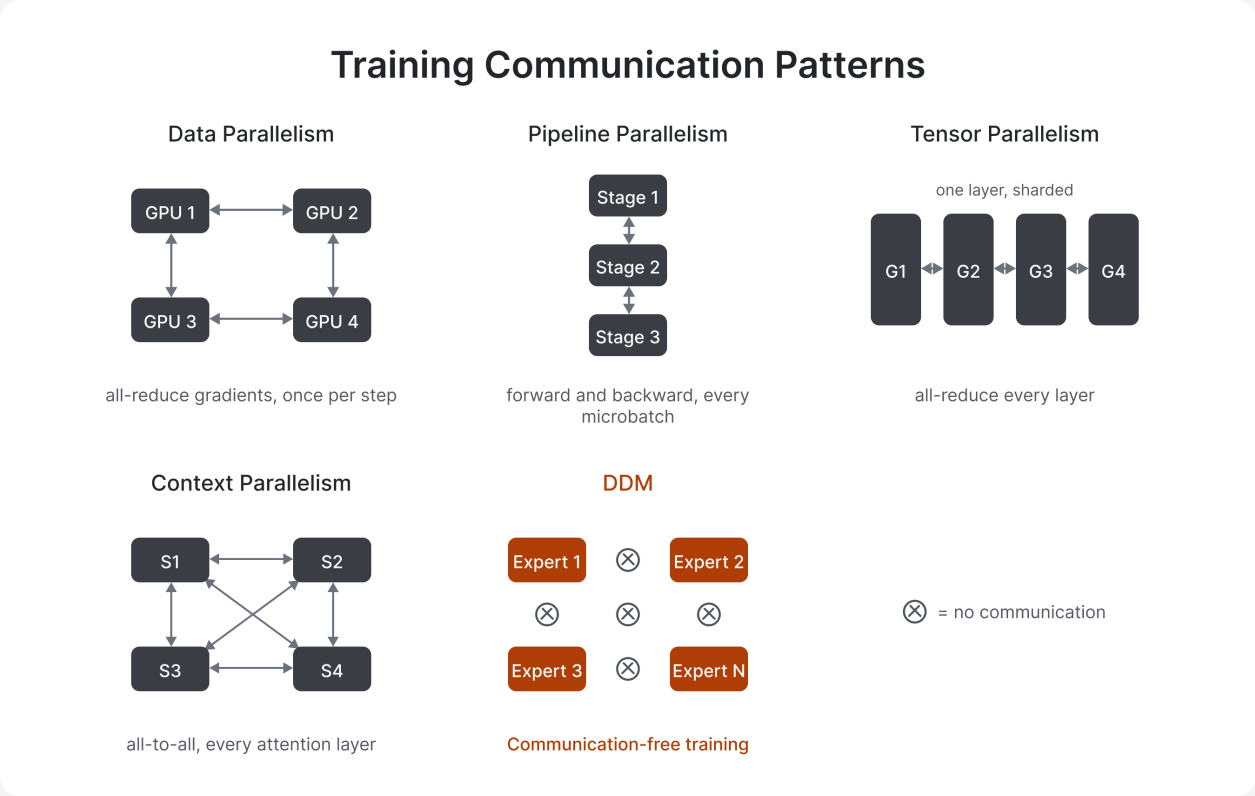

Figure 3 | **Training communication patterns.** Data, pipeline, tensor, and context parallelism each synchronize across GPUs during training, from a gradient all-reduce once per step to an all-to-all exchange at every attention layer. A DDM requires none of it.

3. Method

3.1. Architecture

Paris 2.0 is built from three pieces, a shared preprocessing stack, an expert pool, and a router. Videos are encoded into cached causal HunyuanVAE latents [4], where HunyuanVAE is the video autoencoder that compresses each clip into a compact latent, and prompts are encoded once using T5-v1.1-XXL and CLIP-ViT-L/14. Expert training consumes these cached tensors directly, so the expensive perception stack does not sit in the normal training forward path.

Each expert is an 11B-parameter FLUX-style MM-DiT, the multimodal diffusion transformer backbone that generates the video, initialized only from FLUX.1-dev image weights [1] inside an OpenSora-derived training recipe [9], and it operates on packed video latents.

The router is the lightweight counterpart to the expert pool, a DiT-B model of roughly 100M parameters. It consumes the current noisy video latent, the diffusion timestep, and the pooled 768-dimensional CLIP ViT-L text vector, together with an optional DINOv2 first-frame feature [7], and produces routing weights over the experts. The experts condition on the full CLIP and T5-XXL text embeddings, whereas the router reads only the pooled CLIP vector and never T5. The optional first-frame path, where DINOv2 is a self-supervised vision model that produces general-purpose image features, lets the same router serve both text-to-video and image-to-video generation.

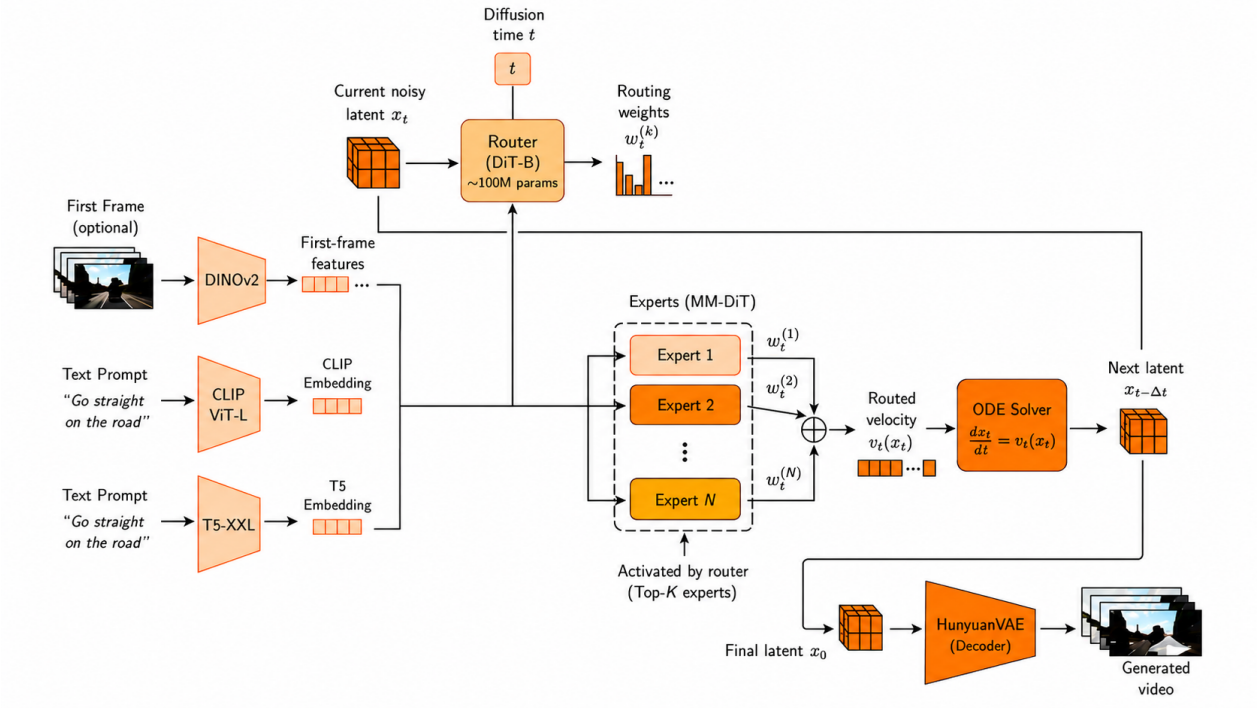

Figure 4 | **The Paris 2.0 inference pipeline.** Text prompts are encoded by T5-XXL and CLIP ViT-L, and an optional first frame by DINOv2. At each denoising step the DiT-B router reads the current noisy latent, the diffusion timestep, and the pooled CLIP vector, with the optional first-frame feature, and produces routing weights that activate the top-K MM-DiT experts. The experts condition on the full CLIP and T5-XXL embeddings, while the router uses only the pooled CLIP vector. The weighted sum of the experts' velocity fields drives an ODE solver to the next latent, and the final latent is decoded to video by HunyuanVAE.

3.2. Training

Each expert is pre-trained from scratch for video generation on a single data cluster, rather than fine-tuned from a pre-existing video model, and is optimized with a flow-matching velocity objective that teaches it to turn noise into video [5]. What matters is not the exact block schedule or checkpoint format but the training boundary, each expert is confined to its own cluster.

The router is trained independently of the experts, as a supervised cluster classifier over noisy latents. From a clean cached latent perturbed with noise at a sampled diffusion timestep under the training noise schedule, with the paired CLIP text feature, the Stage 1 router predicts the source-cluster label. This objective is deliberately simple, and later router stages can use image conditioning and stronger timestep-aware objectives.

4. Experiments

4.1. Setup

For the Stage 1 study reported here, we use three selected clusters and compare two ways of using the same data, a monolithic 11B FLUX-style MM-DiT trained on the union of the clusters, and three 11B experts trained separately, one per cluster. Both paths use the same cached HunyuanVAE/text-embedding pipeline, the same model family, and an aligned flow-matching training recipe. The

comparison is iso-FLOP and iso-data, each expert sees one third of the data with one third of the monolithic FLOPs, leaving parameter count as the only structural difference between them. Our prior Bagel Labs analysis of decentralized diffusion at the routing level [8] shows that this difference in parameter count is not what drives DDM quality, since full-ensemble routing across all experts performs strictly worse than sparse routing despite using maximal capacity.

We evaluate both arms on a deterministic cluster-stratified subset of N=2048 held-out clips at 256×256, using Euler-50 sampling, classifier-free guidance scale 7.5, and a fixed seed, so the two models are scored under an identical generation protocol.

4.2. Results

Table 1 | **Low-resolution text-to-video, compute-matched comparison.** Metrics are computed on the same N=2048 cluster-stratified subset. Arrows indicate preferred direction.

Metric	DDM	Monolithic
FVD ↓	**279.01**	561.04
CLIP text-video ↑	**0.2178 ± 0.0012**	0.2032 ± 0.0011
Aesthetic ↑	**3.9036 ± 0.0082**	3.7950 ± 0.0077
Motion (px/frame)	0.712 ± 0.057	0.555 ± 0.043

As Figure 2 plots in relative terms, the DDM cuts FVD, the measure of distributional realism, from 561.04 to 279.01, and lifts CLIP text-video prompt alignment and aesthetic quality. Motion is a descriptive measure of per-frame displacement magnitude, not of temporal consistency, and carries no preferred direction. The headline result is that the DDM improves distributional and prompt-alignment metrics under the same generation protocol.

4.3. Ablations

The comparison above tests the DDM against the monolithic baseline. To probe whether the gain is genuinely about routing across multiple experts rather than ensembling per se, we run two additional ablations on the same Paris 2.0 expert checkpoints.

Switching schedule. With the router bypassed and manual denoising schedules forced over two experts, an alternating schedule across denoising steps outperforms either single expert on CLIP, aesthetic, and warping simultaneously, and 24 of 40 prompts prefer a switching schedule over the best single expert, as Figure 5 shows. The asymmetry between expert A then expert B and the reverse order is large enough to read as directional specialization across denoising time, one expert prefers high-noise steps and the other prefers low-noise steps. This is the most direct video-scale evidence that the multi-expert routing principle, not raw parameter count, is what carries the headline gain.

Expert specialization. A second probe asks whether each expert actually learns its assigned data cluster, or whether it collapses to the marginal distribution. Evaluating one expert checkpoint on its own cluster's prompts gives CLIP 0.2175, against 0.1781 on a generic prompt set on the same checkpoint, a 0.039 in- versus out-of-cluster gap. Experts genuinely specialize, which is the precondition under which a router can extract value at inference.

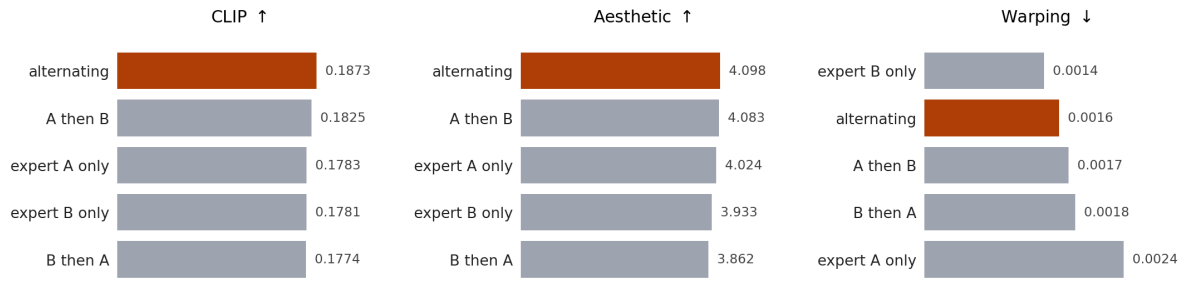

Figure 5 | **Switching schedule ablation.** Manual schedules over two experts with the router bypassed, N=40 video prompts. The alternating schedule wins on all three metrics simultaneously.

5. Discussion

These results show that the DDM recipe survives the move from images to video, where experts specialized on separate data clusters beat the monolithic backbone, extending our prior result that the same holds across heterogeneous training objectives [3].

The same recipe extends naturally to physical AI. Foundation world models built on video diffusion backbones must cover far more environments than any single training cluster can hold, and specializing experts per environment and routing across them at inference fits this regime directly. Scaling along this axis points toward world models trained on the distributed compute the world already has rather than the monolithic clusters only a few labs can assemble.

References

[1] Black Forest Labs. FLUX, 2024. https://github.com/black-forest-labs/flux.

[2] Jiang, Z., Seraj, R., Villagra, M., and Roy, B. Paris: A decentralized trained open-weight diffusion model. *arXiv preprint arXiv:2510.03434*, 2025.

[3] Jiang, Z., Seraj, R., Villagra, M., and Roy, B. Heterogeneous decentralized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[4] Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., and Zhong, C. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[5] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[6] McAllister, D., Tancik, M., Song, J., and Kanazawa, A. Decentralized diffusion models. *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23323–23333, 2025.

[7] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[8] Villagra, M., Roy, B., Seraj, R., and Jiang, Z. Expert-data alignment governs generation quality in decentralized diffusion models. In *ICLR 2026 DeLTa Workshop*, 2026. arXiv:2602.02685.

[9] Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., and You, Y. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.