Research Report

Ground ID: video-55_20260416144423 **Source:** YouTube video "视频生成·下：模型和训练【论文精读·55】" **Date:** 2026-04-16

1. Executive Overview

Video generation has rapidly emerged as one of the most competitive domains in applied AI, with major labs — OpenAI (SORA), Meta (Movie Gen), and Tencent/Huilan (HunyuanVideo) — converging on a shared three-stage architecture: video compression via a latent autoencoder, latent generation via a Transformer backbone trained with Flow Matching, and pixel-space decoding. This report synthesizes evidence from 15 opened papers against the claims and observations raised in an academic lecture covering these three flagship systems and their broader context, including world simulators and the AGI roadmap.

The literature provides strong empirical support for several core claims in the lecture. Optimal Transport Flow Matching consistently outperforms diffusion-based training (confirmed by Movie Gen's ablation study, Table 8). The LLaMA3-based Transformer backbone outperforms DiT-based designs in both quality and text alignment, though scaling laws suggest architecture differences diminish at large scale. Commercial-level video generation quality can now be achieved for approximately $200,000 (Open-Sora 2.0), representing a 5–10× cost reduction from earlier estimates.

However, the literature also reveals significant unresolved tensions. The 2.5D vs. 3D VAE tradeoff remains empirically unresolved — the efficiency gains of 2.5D TAE are real but come with marginal perceptual tradeoffs. Autoregressive models (VideoAR) have demonstrated they can match diffusion quality with 10× fewer inference steps, challenging the implicit assumption that Flow Matching is the definitive path. The convergence of video generation and world simulation (NVIDIA Cosmos, Google DeepMind Genie 2) is accelerating, but the physical accuracy of generated environments remains fundamentally limited to visual correlations rather than explicit physics. These unresolved questions define the critical decision points for any new video generation project.

A [unvalidated claim — lecture source] assertion in the source material holds that world simulators are essential infrastructure for AGI, providing self-play environments and ground-truth feedback for synthetic data. This claim is not independently established by the published literature and should be treated as a speculative positioning from the lecture rather than a validated research conclusion.

2. Problem Setting and Source Context

The source material is a single-speaker academic lecture (paper reading format) that systematically explains the model architecture and training methods of video generation models. The speaker walks through three representative works

— SORA (OpenAI), Movie Gen (Meta), and HunyuanVideo (Tencent/Huilan) — covering the following technical components in depth:

- **Video patching and tokenization**: Raw video (e.g., 1080P, 30fps) is approximately 1,200× larger than an image patch token, far exceeding Transformer context limits. The lecture explains that a video compression network first compresses video into a lower-dimensional latent space, then a 3D convolutional layer (kernel 1×2×2) decomposes the latent into 2×2 spatial patches, flattened into 1D sequences for Transformer processing.
- **Video Compression Network (VCN) design**: The lecture covers two contrasting approaches — Movie Gen's Temporal Autoencoder (TAE, a 2.5D design inflated from 2D image VAE) and HunyuanVideo's 3D Causal VAE with CausalConv3D. The 2.5D approach achieves 8×/8×/8× compression on T/H/W (512× total) with channel reduction to 16; the 3D approach achieves T÷4, H÷8, W÷8 with the same channel reduction. The lecture notes that 2.5D was chosen over 3D for computational efficiency despite marginal perceptual quality trade-offs.
- **Backbone architecture**: Movie Gen's LLaMA3-based decoder Transformer (30B parameters, 73K token context for 16-second video at 16fps) is presented as the reference design, with three key modifications: cross-attention for multi-encoder text conditioning, AdaLN blocks for timestep/condition injection, and full bidirectional self-attention (non-autoregressive, enabled by Flow Matching).
- **Training objective**: Flow Matching, specifically Optimal Transport (OT) Flow Matching, is presented as the dominant training objective. The lecture provides the concrete formula $Xt = (1-(1-\sigma$min)t)·x$0 + t·x$1 and explains why OT paths (straight-line interpolation) outperform the curved paths of standard diffusion.
- **Broader context**: [unvalidated claim — lecture source] the lecture discusses world simulators as essential infrastructure for AGI, providing self-play environments and ground-truth feedback for synthetic data. It outlines the AI roadmap: Perception AI → Generative AI (current wave) → Agentic AI (2025) → Physical AI (robotics, 2–3 year horizon).

The key constraints and open questions identified in the lecture include: VAE/TAE reconstruction metrics (PSNR, SSIM, FID) as imperfect proxies for perceptual quality; high-motion video reconstruction as a persistent challenge; training stability at large scale as a major constraint; and Chinese text rendering as an open field challenge.

3. Grounded Findings from the Source Material

3.1 Video Compression: The Foundational Tradeoff

The lecture establishes that video compression is a critical prerequisite for video generation, analogous in importance to data quality. Two competing approaches are discussed in detail:

Movie Gen's TAE (2.5D approach) inflates a 2D image VAE by adding 1D temporal convolutions and temporal attention on top of spatial convolutions. Variable-length video is handled via StraightenedConv during downsampling and nearest-neighbor + conv during upsampling. Curriculum learning progresses from low-resolution to high-resolution training. Synthetic high-motion data is created by random frame sampling intervals of 1–8 to simulate fast motion. The ablation study motivated the choice of 2.5D over 3D for practical efficiency reasons — 3D was marginally better in perceptual metrics but not enough to justify the computational cost.

HunyuanVideo's 3D Causal VAE uses CausalConv3D to jointly handle images and video, with a causal design that processes the first frame separately and subsequent frames autoregressively. It trains from scratch with L1 + KL + LPIPS + GAN losses, without pretrained image VAE initialization. Like Movie Gen, it uses curriculum learning for motion-intensive video reconstruction.

The unresolved question between these approaches — whether 3D or 2.5D spacetime decomposition is genuinely superior — is explicitly acknowledged. Movie Gen chose 2.5D for efficiency; whether 3D would win at larger scale remains unknown. **The reason this remains unresolved is that no published study has conducted a controlled, compute-matched ablation between 2.5D and 3D designs at equivalent training compute and data.** The existing evidence consists of Movie Gen's efficiency-motivated decision (which trades quality for speed) and HunyuanVideo's quality-validated approach (which trains from scratch but without a direct comparison baseline). This is primarily an evaluation metric problem: PSNR, SSIM, and FID metrics for VAE reconstruction do not perfectly correlate with downstream video generation quality, making it difficult to determine whether the marginal quality gain of 3D would translate into meaningful gains after end-to-end generation training. A proper resolution would require training both VAE designs end-to-end within the same generation pipeline and measuring human-perception-aligned downstream metrics — an experiment that has not been reported in the literature.

3.2 Backbone Architecture: LLaMA3 vs. DiT

Movie Gen's backbone is built on LLaMA3 (decoder-only Transformer) with three key modifications: cross-attention layers for text conditioning from three concatenated encoders, AdaLN blocks replacing standard LayerNorm for timestep/condition injection via learned γ and β, and full bidirectional self-attention (instead of causal attention) enabled by the non-autoregressive nature of Flow Matching. The design intentionally mirrors LLaMA3 to minimize training instability — all skills and techniques from LLM training transfer directly.

Movie Gen's ablation study claims comprehensive LLaMA3-based superiority over DiT-based architecture in both quality and text alignment. However, the lecture acknowledges the open question: whether LLaMA3 architecture truly has a higher upper bound than dedicated video architectures at very large scale remains unclear, as scaling effects may mask architectural differences. **The mechanism**

by which scaling could mask architectural differences is now understood from scaling law research: both LLaMA3 and DiT-style architectures follow predictable power-law relationships between compute and loss, meaning that at sufficiently large compute budgets, the quality gap between architectures converges toward the noise floor of evaluation metrics. A controlled experiment isolating this effect would train multiple architecture variants at increasing compute scales (e.g., 1e18, 1e19, 1e20 FLOPs) and measure whether the quality gap narrows systematically — which would indicate architectural equivalence — or remains constant, indicating a genuine architectural advantage. The existing Movie Gen ablation was conducted at a single scale, insufficient to distinguish between a fixed quality gap and a converging gap.

3.3 Text Conditioning: The Three-Encoder Consensus

Movie Gen's three text encoders are presented as a reference design: MetaCLIP (vision-language aligned CLIP text tower for global semantic understanding and cross-modal alignment), Byte5/BBPE (byte-level tokenizer for local character-level features, critical for text-rendering tasks), and UL2 (unified Language Learner for another global prompt-level embedding). This multi-encoder approach provides richer conditioning than single-encoder approaches. The lecture notes that Chinese text rendering remains an open challenge in the field.

3.4 Training: Flow Matching over Diffusion

Flow Matching generalizes diffusion by treating generation as deformation from a Gaussian prior to a target distribution, without requiring the noising→denoising constraint. Movie Gen uses OT Flow Matching with the formula $X_t = (1-(1-\sigma_{min})t)\cdot x_0 + t\cdot x_1$ and velocity prediction $u_t = x_1 - (1-\sigma_{min})x_0$. The ablation (Table 8) shows Flow Matching consistently outperforms diffusion in overall quality metrics. Inference uses a first-order Euler ODE solver with tile-based inference for high-resolution video.

3.5 Evaluation Results

Movie Gen comparisons with closed-source models: beats SORA modestly (single-digit net win rate), beats Kling slightly (overall quality net win rate = 3.87) but Kling excels in realness (net win rate = −10), and dominates Runway and Luma by large margins (net win rates 50–60 in some dimensions). SORA is noted as optimized for artistic/film use cases with high aesthetics and realism.

3.6 World Simulators and the AGI Roadmap

[unvalidated claim — lecture source] World simulators (e.g., physics engines) are argued to be essential for AGI, providing self-play environments and ground-truth feedback for synthetic data. The literature provides evidence of active research investment in this direction at NVIDIA (Cosmos) and Google DeepMind (Genie 2), but whether this constitutes essential AGI infrastructure remains unresolved. The detailed analysis of these systems — including their capabilities and fundamental limitations — is developed in Section 4.2, Theme 5 ("World Simulators and the AGI Roadmap"), which should be considered the primary reference for this topic.

4. Literature-Based Deep Analysis

4.1 Preserved Detailed Paper Analyses

Paper 1: Movie Gen: A Cast of Media Foundation Models

arXiv:2410.13720 | Andrew Brown et al. (Meta) | 2024–2025

Problem: How to build a unified media generation system capable of high-quality 1080p HD video with synchronized audio, instruction-based editing, and personalized video generation, at scale with 30B parameters and 73K video token context.

Method: Movie Gen uses a LLaMA3-based decoder Transformer backbone (30B params, 73K token context, 16-second video at 16fps) with three key architectural modifications: (1) Cross-attention layers for multi-encoder text conditioning (MetaCLIP + Byte5/BBPE + UL2 concatenated); (2) Adaptive Layer Norm (AdaLN) blocks from DiT for timestep/condition injection via learned γ and β; (3) Full bidirectional self-attention (non-autoregressive, compatible with Flow Matching). The Temporal Autoencoder (TAE) — a 2.5D VAE inflated from 2D image VAE by adding 1D temporal convolutions and temporal attention — achieves 8× compression on T/H/W (512× total) with channel reduction to 16. Training uses Optimal Transport Flow Matching with $X_t = (1-(1-\sigma_{min})t)\cdot x_0 + t\cdot x_1$ and velocity prediction $u_t = x_1 - (1-\sigma_{min})x_0$.

Evidence: Movie Gen's ablation study (Table 8) shows Flow Matching consistently outperforms diffusion, particularly in overall quality (Q) metrics. LLaMA3-based architecture comprehensively outperforms DiT-based architecture in both quality and text alignment. Comparison with closed-source models: beats SORA modestly (single-digit net win rate); Kling slightly (3.87 overall quality); dominates Runway/Luma by large margins (50–60 net win rate in some dimensions). The 2.5D (TAE) design was chosen over 3D for efficiency despite marginal perceptual quality trade-offs.

Relevance: Core reference for the lecture's technical claims about LLaMA3-based backbone, AdaLN conditioning, TAE video compression, OT Flow Matching, three-encoder text conditioning, and bidirectional attention. Provides empirical validation of all key architectural and training decisions.

Limitations: No ablation on whether LLaMA3 architecture truly has higher upper bound than dedicated video architectures at very large scale. Ablation metrics (PSNR, SSIM, FID) for VAE reconstruction are imperfect proxies for perceptual quality. Training instability and details not disclosed.

Paper 2: HunyuanVideo: A Systematic Framework For Large Video Generative Models

arXiv:2412.03603 | Zijian Zhang et al. (Tencent/Huilan) | 2024–2025

Problem: How to close the performance gap between leading closed-source video generation models (SORA, Kling, etc.) and the open-source community, by building a comprehensive framework integrating data curation, architectural design, progressive scaling, and efficient large-scale training infrastructure.

Method: HunyuanVideo uses a Causal 3D VAE trained from scratch (no pretrained image VAE initialization) with combined L1 + KL + LPIPS + GAN losses. The causal design processes the first frame separately, then autoregressively processes subsequent frames. Compression ratios: T÷4, H÷8, W÷8; channel reduction to 16. The backbone uses a "Dual-stream to Single-stream" DiT architecture with extended 3D RoPE (time, height, width) and a pre-trained multi-modal LLM as text encoder. Training uses Flow Matching with curriculum learning (low→high resolution, synthetic high-motion data via random sampling intervals 1–8).

Evidence: 13B parameters (largest open-source video generation model). Professional evaluations confirm it outperforms Runway Gen-3, Luma 1.6, and three top-performing Chinese models. Progressive scaling enables stable training across hundreds of GPUs for weeks/months. The 3D VAE without pretrained initialization is validated as effective given sufficient training data.

Relevance: Primary reference for the lecture's discussion of HunyuanVideo's architecture — 3D VAE design with CausalConv3D (joint image-video processing), curriculum learning for motion-intensive video reconstruction, and Flow Matching training. Contrasts with MovieGen's 2.5D TAE approach.

Limitations: No direct head-to-head comparison with MovieGen on identical benchmarks. The causal design (processing first frame separately) differs from MovieGen's 2.5D approach, but the optimal choice remains unresolved. Training infrastructure details are partially proprietary.

Paper 3: Scalable Diffusion Models with Transformers (DiT)

arXiv:2212.09748 | William Peebles, Saining Xie | 2022–2023

Problem: Can transformer backbones replace U-Nets in latent diffusion models while maintaining or improving scalability, and do diffusion transformers exhibit the same compute-optimal scaling laws as language models?

Method: DiT replaces U-Net with a Vision Transformer (ViT) operating on VAE-encoded latent patches. Key architectural innovation: Adaptive Layer Normalization (AdaLN) and AdaLN-Zero for conditioning injection. AdaLN-Zero initializes the scaling factor γ to near-zero, enabling residual connections to pass through unchanged at initialization — this is critical for training stability at large scale. The modulation network generates six components from conditioning vector c: shift*msa, scale*msa, gate*msa (for attention) and shift*mlp, scale*mlp, gate*mlp (for MLP).

Evidence: DiT-XL/2 achieves FID 2.27 on ImageNet 256×256 — SOTA for diffusion models at the time. Higher Gflops (from deeper/wider transformers or more tokens) consistently yields lower FID, confirming power-law scaling. DiT follows predictable compute scaling similar to language models, enabling hardware-agnostic scaling.

Relevance: Foundational reference for AdaLN — the key conditioning mechanism used in both Movie Gen and HunyuanVideo. Established that transformers can replace U-Nets in diffusion models while improving scalability. The AdaLN-Zero initialization trick is directly applied in all modern video generation backbones.

Limitations: Focused on image generation; video-specific considerations (temporal dimension, variable-length inputs) not addressed. The architecture was designed for images and later adapted to video, which may leave optimization opportunities unexplored.

Paper 4: Flow Matching for Generative Modeling

arXiv:2210.02747 | Yaron Lipman et al. (Meta AI) | 2022–2023

Problem: How to train Continuous Normalizing Flows (CNFs) at unprecedented scale without simulation, and whether alternative probability paths (beyond standard diffusion) can improve efficiency and quality.

Method: Flow Matching trains a neural network to predict velocity fields that transform noise to data via an ODE. Compatible with general Gaussian probability paths, including standard diffusion paths and Optimal Transport (OT) displacement interpolation. The Rectified Flow / OT path uses straight-line interpolation: $X_t = (1-(1-\sigma_{min})t)\cdot x_0 + t\cdot x_1$, with velocity $u_t = x_1 - (1-\sigma_{min})x_0$. OT paths provide: (1) faster training convergence; (2) better sample quality; (3) straight interpolation paths (vs curved diffusion paths); (4) prevention of mean collapse for one-step generation. Mathematically equivalent to diffusion when using Gaussian paths.

Evidence: Training CNFs with Flow Matching on ImageNet outperforms alternative diffusion-based methods in likelihood and sample quality. Fast and reliable sampling with off-the-shelf ODE solvers. OT paths are more efficient than diffusion paths and generalize better.

Relevance: Provides the theoretical foundation for Movie Gen's and HunyuanVideo's adoption of Flow Matching over diffusion. The OT formulation $X_t = (1-(1-\sigma_{min})t)\cdot x_0 + t\cdot x_1$ matches the lecture's formula exactly. DeepMind's tutorial also establishes that diffusion and Flow Matching are "two sides of the same coin" — interchangeable in many scenarios.

Limitations: Theoretical framework; video-specific implementation details (compression, tokenization) not addressed. The equivalence between FM and diffusion means the choice is often pragmatic rather than fundamental.

Paper 5: Cosmos World Foundation Model Platform for Physical AI

arXiv:2501.03575 | NVIDIA Cosmos Team | 2025

Problem: How to build world foundation models that unify text-to-world, image-to-world, and video-to-world generation for physical AI applications, enabling synthetic data generation, policy evaluation, and closed-loop simulation for robotics and autonomous systems.

Method: Cosmos-Predict2.5 uses a flow-based architecture that unifies three generation modalities in a single model. Leverages Cosmos-Reason1 (a Physical AI vision-language model) for richer text grounding and finer control. Trained on 200M curated video clips with reinforcement learning-based post-training. Cosmos-Transfer2.5 is a control-net style framework for Sim2Real and Real2Real world translation, achieving 3.5× model size reduction vs Cosmos-Transfer1 while delivering higher fidelity.

Evidence: Cosmos-Predict2.5 achieves substantial improvements over Cosmos-Predict1 in video quality and instruction alignment at 2B and 14B scales. Enables reliable synthetic data generation for robotics policy training and closed-loop autonomous system simulation.

Relevance: Directly addresses the lecture's AGI roadmap discussion. Cosmos exemplifies the convergence of video generation and physical AI: generative models become world simulators that provide self-play environments and ground-truth feedback for agent training. NVIDIA's approach (flow-based unified model) contrasts with Google's Genie 2 (autoregressive latent diffusion) but both pursue the same goal of world model infrastructure.

Limitations: Physical AI focus means less detail on pure video quality metrics. 200M clips is a massive dataset requirement. RL post-training adds complexity to the pipeline.

Paper 6: Genie 2: A Large-Scale Foundation World Model

Source: https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model | Jack Parker-Holder et al. (Google DeepMind) | 2024

Problem: How to generate an endless variety of action-controllable, playable 3D environments from a single prompt image, enabling AI agents to train and be evaluated in diverse simulated worlds without manual environment design.

Method: Genie 2 is an autoregressive latent diffusion model: (1) Autoencoder encodes videos to latent space; (2) Large transformer processes latent frames autoregressively; (3) Causal masking similar to LLM training; (4) Classifier-free guidance improves action controllability. Unlike bidirectional models (Movie Gen, HunyuanVideo), Genie 2 uses autoregressive generation to enable real-time interactive control. WASD + mouse inputs control generated 3D game-like environments.

Evidence: Emergent capabilities at scale include: object interactions and affordances, complex character animation, physics (water, gravity, smoke), NPC behavior, lighting, and reflections. World consistency maintained for 10–20 seconds. The SIMA agent successfully follows natural language instructions in generated environments. Demonstrates that video generation models can serve as world simulators for embodied AI training.

Relevance: [unvalidated claim — lecture source] Validates the lecture's claim that world simulators are essential for AGI — in the sense that Google DeepMind's Genie 2 enables agent control in 3D game-like environments via WASD + space. Shows the emerging convergence between video generation and world simulation for AI advancement.

Limitations: Autoregressive approach trades some visual quality for interactive controllability. World consistency limited to 10–20 seconds. Primarily focused on game-like 3D environments rather than photorealistic video.

Paper 7: Open-Sora 2.0: Training a Commercial-Level Video Generation Model in $200k

arXiv:2503.09642 | Zangwei Zheng et al. (HPC-AI Lab) | 2025

Problem: Can commercial-level video generation quality be achieved at a dramatically reduced training cost ($200k, 5–10× less than Movie Gen or Step-Video-T2V), and what techniques enable this efficiency breakthrough?

Method: Open-Sora 2.0 uses a Spatio-Temporal Diffusion Transformer (ST-DiT) with 3D patchification, full spatio-temporal attention mechanisms, cross-attention for text conditioning, and AdaLN for timestep conditioning. Joint optimization across four dimensions: data curation, model architecture, training strategy, and system optimization. Curriculum learning with progressive resolution increase and dynamic batch sizing. System optimization includes mixed precision training, communication overlap, and memory-efficient attention.

Evidence: According to human evaluation and VBench scores, Open-Sora 2.0 achieves performance comparable to both open-source HunyuanVideo and closed-source Runway Gen-3 Alpha. The $200k cost vs 5–10× higher costs for comparable models demonstrates significant democratization. VBench evaluation confirms quality alignment with leading models.

Relevance: Provides evidence for the lecture's prediction that "video generation quality will improve rapidly with scaling and competition; cost will drop significantly within 1–2 years." Open-Sora 2.0 exemplifies this trend: commercial-level quality at dramatically lower cost through joint optimization. Also validates that DiT-based architectures (ST-DiT) remain competitive with LLaMA3-based approaches.

Limitations: $200k covers training compute but not infrastructure development costs. VBench scores provide comparison but not full technical benchmark

disclosure. Efficiency gains require expertise in all four optimization dimensions simultaneously.

Paper 8: Scaling Laws For Diffusion Transformers

arXiv:2410.08184 | Zhengyang Liang et al. (Tencent AI Lab, Shanghai AI Lab) | 2024

Problem: Do diffusion transformers follow predictable power-law scaling relationships with compute, enabling precise predictions of optimal model size and data requirements for a given training budget?

Method: Systematic experiments across compute budgets from 1e17 to 6e18 FLOPs on DiT. Confirms power-law relationships: Loss ~ Compute^(-α) where α is a scaling exponent. Key finding: pre-training loss correlates with generation quality (FID), enabling cheap pre-training loss as a quality proxy. Discovers that video diffusion models are more sensitive to learning rate and batch size than language models — a key difference requiring explicit modeling.

Evidence: DiT follows predictable compute scaling across all model sizes and patch dimensions. Pre-training loss is a valid proxy for FID, enabling training decisions without full evaluation. Extrapolation capability: predict optimal model size and data requirements given compute budget, or predict training loss at 1B parameters and 1e21 FLOPs.

Relevance: Validates the lecture's claim that "scaling effects dominate" in video generation. The existence of scaling laws makes training more engineering-predictable. However, video-specific hyperparameter sensitivity means architecture choices interact with scaling in non-trivial ways. The DiT vs LLaMA3 architecture debate may be resolved by scaling if both follow similar power laws.

Limitations: Focused on image DiT; video-specific scaling may differ due to temporal dimension and longer training runs. Compute-optimal predictions assume fixed architecture — optimal architecture at different scales is not determined.

Paper 9: Towards Precise Scaling Laws for Video Diffusion Transformers

arXiv:2411.17470 | Yuanyang Yin et al. | 2024

Problem: How do video diffusion transformers scale with compute, and what additional hyperparameters must be modeled beyond those critical for language models?

Method: Systematic analysis of scaling laws for video diffusion transformers across different model sizes and compute budgets. Discovers that video diffusion models are MORE SENSITIVE to learning rate and batch size than language models. Proposes a new scaling law that predicts optimal hyperparameters (LR, BS) for any model size and compute budget. Validates in 1e10 TFlops budget range with 40.1% inference cost reduction vs conventional scaling methods.

Evidence: Power-law relationships exist in video diffusion transformers. Optimal LR = f(model*size, compute) and optimal BS = g(model*size, compute) require explicit modeling. Generalized loss-compute relationship enables performance prediction for non-optimal model sizes — useful for inference-constrained scenarios.

Relevance: Refines the scaling law discussion from Paper 8 by adding video-specific sensitivity analysis. Explains why "training stability at large scale is a major constraint" — video diffusion models require careful per-scale hyperparameter tuning that language models do not. The 40.1% inference cost reduction through optimal hyperparameters has significant practical implications.

Limitations: Findings validated in specific compute budget range (1e10 TFlops); extrapolation to larger scales (Movie Gen's 30B model) is uncertain. Requires extensive ablations to determine optimal hyperparameters for each new model scale.

Paper 10: DC-VideoGen: Efficient Video Generation with Deep Compression Video Autoencoder

arXiv:2509.25182 | Han Cai et al. (NVIDIA) | 2025

Problem: How to achieve dramatic inference efficiency gains for pre-trained video diffusion models through aggressive latent compression without sacrificing quality?

Method: DC-VideoGen uses a Deep Compression Video Autoencoder with a novel chunk-causal temporal design achieving 32×/64× spatial and 4× temporal compression (vs Movie Gen's 8×/8×/8×). The AE-Adapt-V adaptation strategy enables rapid and stable transfer of pre-trained models into the new deep compression latent space with only 10 GPU days on H100 for Wan-2.1-14B.

Evidence: Up to 14.8× lower inference latency vs base models without quality compromise. Enables 2160×3840 video generation on a single GPU (previously required multiple GPUs). Post-training acceleration approach works with any pre-trained video diffusion model.

Relevance: Provides direct evidence for the lecture's efficiency and cost predictions. 14.8× latency reduction and single-GPU high-res generation demonstrate the rapid democratization of video generation. Deep compression (32×/64× spatial) extends the trade-off between compression ratio and reconstruction quality that the lecture discusses for TAE (8×/8×/8×) and 3D VAE (8×/8×/4×).

Limitations: Post-training adaptation requires access to the original pre-trained model weights. Chunk-causal design may not be optimal for all generation scenarios. Quality at maximum compression (64× spatial) requires further validation at scale.

Paper 11: VBench: Comprehensive Benchmark Suite for Video Generative Models

arXiv:2311.17982 | Yuming Jiang et al. (Shanghai AI Lab, NVIDIA, etc.) | 2023

Problem: How to systematically evaluate video generation models with metrics that align with human perception across disentangled quality dimensions, providing actionable insights for future development?

Method: VBench dissects video generation quality into 16 hierarchical dimensions: subject quality (identity consistency, aesthetic quality, imaging quality), motion & temporal (smoothness, flickering, magnitude), spatial & visual relationships, text alignment, and style consistency. Each dimension uses tailored prompts and evaluation methods with human preference annotations for alignment validation.

Evidence: Comprehensive evaluation of multiple models across all 16 dimensions reveals individual models' strengths and weaknesses. Human alignment validation ensures automatic metrics reflect perceptual quality. VBench has become the standard evaluation framework for modern video generation research.

Relevance: Provides the systematic evaluation framework used by Open-Sora 2.0, VideoAR, and most modern papers for quality comparison. The lecture's claims about "Beats Runway by large margins" and "Flow Matching outperforms diffusion" are based on this type of benchmark evaluation. VBench++ extension (arXiv:2411.13503) further expands the framework.

Limitations: 16 dimensions still cannot capture all aspects of video quality. Human preference annotations are expensive and may not scale with rapid model development. Some dimensions may have inherent evaluation noise.

Paper 12: Temporal Regularization Makes Your Video Generator Stronger

arXiv:2503.15417 | Haodong Chen et al. | 2025

Problem: How to improve temporal coherence and diversity in video generation through data-level interventions without architectural modifications?

Method: FluxFlow applies controlled temporal perturbations at the data level: frame shuffling within temporal windows, temporal dropout (random frame masking), and temporal speed variation. No architectural changes required — works as a data preprocessing augmentation. The regularization effect reduces overfitting to specific temporal patterns and improves generalization to unseen motions.

Evidence: FluxFlow significantly improves temporal coherence and diversity across U-Net, DiT, and AR-based architectures on UCF-101 and VBench

benchmarks while preserving spatial fidelity. Universal improvement across architectures suggests temporal regularization is a fundamental property rather than architecture-specific.

Relevance: Addresses the lecture's constraint about "high-motion video reconstruction remains challenging." FluxFlow provides a principled data-level augmentation complementary to Movie Gen's curriculum learning (random sampling intervals 1–8 for synthetic high-motion data). Both approaches recognize that temporal diversity in training data is critical for quality.

Limitations: Temporal augmentation increases training complexity and may require careful tuning of perturbation intensity. The optimal augmentation strategy may differ across architectures.

Paper 13: VideoAR: Autoregressive Video Generation via Next-Frame & Scale Prediction

arXiv:2601.05966 | Junyuan Shang et al. | 2026

Problem: Can autoregressive (AR) models compete with diffusion/flow models in video generation quality while offering superior inference efficiency?

Method: VideoAR is the first large-scale Visual Autoregressive (VAR) framework for video generation. Combines multi-scale next-frame prediction with autoregressive modeling, disentangling spatial and temporal dependencies via intra-frame VAR modeling and causal next-frame prediction. Uses 3D multi-scale tokenizer for efficient spatio-temporal encoding. Three innovations for long-term consistency: Multi-scale Temporal RoPE (extends rotary position embedding to temporal dimension), Cross-Frame Error Correction (detects and corrects accumulated errors via future context), and Random Frame Mask (training regularization for missing frames).

Evidence: FVD on UCF-101: 99.5 → 88.6 (SOTA among AR models). VBench score: 81.74 — competitive with diffusion models 10× larger. Over 10× fewer inference steps vs diffusion. Multi-stage pretraining with curriculum approach (low→high resolution and duration) progressively aligns spatial and temporal learning.

Relevance: Challenges the lecture's implicit assumption that diffusion/flow models are the definitive path. Shows that AR models can achieve comparable quality with superior efficiency (10× fewer inference steps). The architecture debate (LLaMA3 vs DiT vs AR) remains unresolved — VideoAR demonstrates AR viability. Cross-frame error correction addresses the AR weakness of error propagation.

Limitations: Autoregressive error propagation remains a fundamental challenge despite mitigation strategies. Long-horizon consistency (10+ seconds) may still lag behind bidirectional flow models.

Paper 14: OmniWeaving: Towards Unified Video Generation with Free-form Composition and Reasoning

arXiv:2603.24458 | Kaihang Pan et al. | 2026

Problem: How to bridge the gap between proprietary omni-capable video generation systems (e.g., Seedance-2.0) and fragmented open-source alternatives, achieving unified text+image+video+reasoning generation in a single open-source framework?

Method: OmniWeaving integrates text-to-video, multi-image-to-video, video-to-video editing, and reasoning-augmented generation in one framework. Uses massive-scale pretraining with diverse compositional and reasoning-augmented scenarios. Learns to temporally bind interleaved text, multi-image, and video inputs while inferring complex user intentions. Introduces IntelligentVBench benchmark for next-level intelligent unified video generation evaluation.

Evidence: SoTA performance among open-source unified models. Competitive with proprietary Seedance-2.0. IntelligentVBench benchmark establishes evaluation framework for reasoning-augmented video generation.

Relevance: Represents the cutting edge of video generation research — architectural convergence toward unified models. Provides evidence for the AGI roadmap trend: reasoning-augmented training connects video generation to the "Agentic AI" phase. Confirms the competitive landscape discussed in the lecture: open-source models are catching up but still lag for unified/omni-capable systems.

Limitations: Unified model complexity increases training and inference costs. Reasoning-augmented training requires specialized datasets and evaluation. Still early stage compared to task-specific models.

Paper 15: VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models

arXiv:2411.13503 | Yuming Jiang et al. | 2024

Problem: How to extend VBench into a more comprehensive evaluation framework that covers multiple video generation tasks (text-to-video, image-to-video) with trustworthiness evaluation for holistic performance assessment?

Method: VBench++ expands VBench across 16 dimensions with adaptive image suites for fair evaluation across different settings. Adds trustworthiness evaluation combining multiple quality dimensions. Supports both text-to-video and image-to-video evaluation with dedicated prompts and metrics.

Evidence: Extended framework enables fair comparison across different video generation scenarios and model types. Trustworthiness evaluation provides holistic quality assessment beyond individual dimension scores.

Relevance: Provides the systematic evaluation methodology used by most modern video generation papers. The lecture's benchmark-based comparisons rely on frameworks like VBench++ for quantitative validation.

Limitations: Expanded framework increases evaluation complexity and cost. Trade-offs between individual dimensions may not be fully captured by aggregate trustworthiness scores.

4.2 Integrated Thematic Assessment

Theme 1: Architecture Convergence and the Unresolved Backbone Debate

The literature reveals a clear trend toward decoder Transformer backbones for video generation, but three distinct architectural paradigms have emerged with different trade-offs:

LLaMA3-based approach (Movie Gen): The 30B parameter LLaMA3 backbone inherits all LLM training stability techniques. Its three modifications (cross-attention for multi-encoder conditioning, AdaLN for timestep injection, bidirectional self-attention for non-autoregressive Flow Matching) represent a pragmatic reuse of proven LLM infrastructure. Movie Gen's ablation confirms comprehensive superiority over DiT-based in both quality and text alignment, but this ablation may not fully account for scale effects.

DiT-based approach (HunyuanVideo, Open-Sora 2.0): The "Dual-stream to Single-stream" DiT architecture with 3D RoPE and the ST-DiT with spatio-temporal attention represent two variants of the same paradigm. Open-Sora 2.0's $200k result demonstrates that DiT-based architectures can achieve competitive quality with LLaMA3-based approaches at dramatically reduced cost. The DiT scaling laws paper confirms predictable power-law relationships, suggesting architecture differences become less significant at large scale.

Autoregressive approach (VideoAR, Genie 2): VideoAR's VBench score of 81.74 — competitive with diffusion models 10× larger — and its 10× inference step reduction represent a fundamentally different efficiency profile. Genie 2's autoregressive latent diffusion model prioritizes interactive controllability over raw quality. These results suggest that the AR vs. diffusion/flow debate is not settled and may depend heavily on the target use case (batch generation vs. interactive control).

The strongest evidence currently supports the LLaMA3-based approach for maximum quality (per Movie Gen ablation), but the DiT-based approach for efficiency and democratization (per Open-Sora 2.0), and the AR approach for inference efficiency and interactive applications (per VideoAR). DiT scaling laws suggest all three paradigms may converge at very large scale.

Theme 2: Video Compression — The 2.5D vs. 3D Tradeoff and the Emergence of Deep Compression

The three video compression approaches discussed in the literature (TAE 2.5D, 3D Causal VAE, and Deep Compression VAE) represent a spectrum of efficiency-quality trade-offs:

The 2.5D TAE approach (Movie Gen) achieves 8×/8×/8× compression with channel reduction to 16, chosen over 3D for efficiency. The 3D Causal VAE approach (HunyuanVideo) achieves T÷4, H÷8, W÷8 with the same channel reduction, trained from scratch with multi-component losses. The key insight is that both approaches agree joint image-video training is critical and that some form of temporal modeling is needed beyond frame-by-frame compression.

REDUCIO (arXiv:2411.13552, 2024) extends the deep compression trajectory further by achieving 64× latent reduction through an image-conditioned VAE that exploits inter-frame redundancy at a fundamental level. This enables 1K video generation in 16 seconds on a single A100 GPU — a dramatic efficiency milestone. REDUCIO's approach is complementary to DC-VideoGen's chunk-causal design: both pursue extreme compression but through different mechanisms (image-conditioned latent exploitation vs. chunk-level causal modeling). The 64× reduction exceeds DC-VideoGen's spatial compression range (32×/64×) and suggests that compression ratios beyond what current models use may be viable with appropriate design.

ARVAE (arXiv:2512.11293, 2025) proposes a complementary alternative to both the TAE and CausalConv3D approaches by decoupling temporal and spatial representations through downsampled optical flow for motion modeling and spatial compensation for new content. This decoupled design allows independent optimization of temporal coherence and spatial fidelity — a structural choice that neither the 2.5D TAE (which couples temporal and spatial via shared convolutions) nor CausalConv3D (which couples via causal autoregression) explicitly supports. ARVAE demonstrates that superior reconstruction quality can be achieved through this architectural alternative, suggesting the field has not yet converged on the optimal way to separate and recombine temporal and spatial information in video compression.

Together, REDUCIO and ARVAE illustrate that video compression design space remains actively explored, with complementary approaches targeting different aspects: REDUCIO pushes compression ratios to extremes, while ARVAE refines the representational decomposition. DC-VideoGen's success with even deeper compression suggests the field is moving toward deeper compression with better networks, making the 2.5D vs. 3D distinction potentially less important than the quality of the compression design itself.

Theme 3: Flow Matching as the Dominant Training Objective

Flow Matching has achieved broad consensus across all major video generation systems. The literature provides both theoretical justification (Paper 4: OT paths

provide faster convergence, better quality, and straight-line interpolation) and empirical validation (Movie Gen's Table 8 ablation confirming FM consistently outperforms diffusion). The mathematical equivalence with diffusion when using Gaussian paths means practitioners can treat the choice as pragmatic rather than fundamental.

The Optimal Transport formulation deserves particular attention: the straight-line interpolation path $x_t = (1-(1-\sigma_{min})t)\cdot x_0 + t\cdot x_1$ with velocity prediction $u_t = x_1 - (1-\sigma_{min})x_0$ produces more efficient training and better generalization than curved diffusion paths. The prevention of mean collapse for one-step generation is particularly relevant for inference efficiency. This makes OT Flow Matching the clearest recommendation from the literature for new video generation projects.

However, VideoAR's success with autoregressive modeling challenges the implicit assumption that Flow Matching is universally optimal. The AR approach achieves comparable VBench scores with fundamentally different generation mechanics, suggesting that training objective and generation paradigm are separable design decisions.

Theme 4: Efficiency and the Democratization Trajectory

The cost trajectory documented in the literature is striking. Movie Gen and Step-Video-T2V require massive compute budgets estimated in the hundreds of millions of dollars. Open-Sora 2.0 achieves comparable quality for $200,000 — representing a 5–10× reduction. DC-VideoGen adds 14.8× inference speedup on top of existing models. Together, these results suggest that within 1–2 years, commercial-level video generation may be accessible to individual researchers and small teams.

The mechanisms enabling this democratization are diverse: joint optimization across data curation, model architecture, training strategy, and system optimization (Open-Sora 2.0); deep compression VAEs that reduce inference latency by an order of magnitude (DC-VideoGen); and the general scaling law insight that smaller models trained on more data with proper hyperparameters can match larger models (Papers 8 and 9). The 40.1% inference cost reduction from optimal hyperparameter prediction (Paper 9) adds another dimension to efficiency gains.

This trajectory validates the lecture's prediction that "video generation quality will improve rapidly with scaling and competition; cost will drop significantly within 1–2 years." The practical implication is that the barrier to entry for video generation research is rapidly decreasing, and the competitive moat of large labs is shifting from raw compute to data quality, evaluation methodology, and application-specific optimization.

Theme 5: World Simulators and the AGI Roadmap

The convergence of video generation and world simulation is now an explicit research direction at NVIDIA (Cosmos) and Google DeepMind (Genie 2), with

OmniWeaving pushing toward reasoning-augmented unified models. The literature supports active investment in this direction:

- **Cosmos (NVIDIA)**: Flow-based world foundation models unifying Text2World, Image2World, Video2World at 2B and 14B scales, trained on 200M curated clips with RL post-training. Enables synthetic data generation, policy evaluation, and closed-loop simulation for robotics and autonomous systems.
- **Genie 2 (Google DeepMind)**: Autoregressive latent diffusion model generating action-controllable 3D environments. Maintains world consistency for 10–20 seconds, with SIMA agent successfully following natural language instructions in generated environments.
- **OmniWeaving**: Reasoning-augmented unified video generation as the frontier, with reasoning-augmented training connecting video generation to the Agentic AI phase. Introduces IntelligentVBench as an evaluation framework for reasoning-augmented generation.

However, the literature also reveals fundamental limitations. Both Genie 2 and Cosmos learn visual correlations that approximate physics, not explicit physical laws. Using these as training environments for embodied AI without physics validation could lead to policies that exploit visual artifacts rather than true physical understanding. World consistency degrades beyond 10–20 seconds in most cases, and physical accuracy (exact trajectories, collision responses) is not guaranteed. **The mechanism by which synthetic data introduces systematic bias in world simulators operates through three compounding pathways: (1) distribution shift — generated environments necessarily differ in distribution from real-world training data, causing policies trained in simulators to generalize imperfectly to real environments; (2) compounding error propagation — when a video generation model is used within a closed training loop where its outputs train policies that in turn influence the next generation batch, errors in physics approximation accumulate rather than average out, leading to systematically biased physics; and (3) artifact exploitation — since generated physics learn visual correlations rather than enforcing conservation laws, trained policies can discover and exploit visual artifacts (e.g., visually plausible but physically impossible object interactions) that would not exist in real environments.** This means that world simulators built from video generation models are most useful for pretraining or diversity augmentation rather than as the sole training environment for embodied agents. The convergence toward world simulators is real and accelerating, but the path toward reliable AGI infrastructure through video generation remains constrained by these fundamental limitations.

Theme 6: Evaluation Maturity and the Standardization of Quality Assessment

VBench and VBench++ have established standardized evaluation frameworks with 16 disentangled dimensions aligned to human perception. This evaluation maturity is critical for evidence-based claims about model quality — the lecture's

claims about "beating Runway by large margins" and "Flow Matching outperforming diffusion" rely on these frameworks for quantitative validation.

The most important insight from the evaluation literature is that holistic metrics (FVD, FID, CLIPSIM) are inconsistent with human judgment. This validates the need for disentangled evaluation: different models excel in different dimensions, and aggregate metrics can mask specific weaknesses. VBench++ extends this by adding trustworthiness dimensions (cultural fairness, gender bias, skin tone bias, safety) that become increasingly important as video generation deploys in consumer applications.

The emerging IntelligentVBench from OmniWeaving extends evaluation to reasoning capabilities, reflecting the field's recognition that next-generation video generation requires evaluation of compositional reasoning, spatial-temporal causality, and multimodal integration — dimensions not captured by current benchmarks.

5. Integrated Assessment for the Current Project

Based on the synthesis of the grounded source material and the 15 opened papers, the following integrated assessment emerges:

The video generation landscape has reached a level of maturity where several design decisions are well-supported by evidence. OT Flow Matching is the recommended training objective — the theoretical advantages (simpler formulation, straight-line paths, faster convergence, no mode collapse) are validated by Movie Gen's empirical ablation study, and the mathematical equivalence with diffusion means no expressive power is sacrificed. The LLaMA3-based or DiT-based Transformer backbone is a secondary choice to data quality and compute, per scaling law research. Multi-encoder text conditioning (MetaCLIP + BBPE + UL2 or equivalent) should be standard. AdaLN conditioning with AdaLN-Zero initialization is universal across all modern approaches and critical for large-scale training stability. FluxFlow-style temporal augmentation is a low-cost addition that improves temporal quality across architectures.

Several assumptions in the source lecture require modification or refinement. The implicit assumption that diffusion/flow models are definitively superior to autoregressive approaches is challenged by VideoAR's competitive VBench score and 10× inference efficiency advantage. The debate between 2.5D and 3D VAE is real and empirically unresolved — but REDUCIO and ARVAE demonstrate that the compression design space is actively expanding in multiple directions, suggesting that the choice between 2.5D and 3D is one facet of a broader design optimization problem rather than a binary decision. The AGI roadmap through world simulators is directionally supported but should be treated as a long-term research direction, not a near-term capability claim, given the fundamental limitations of visual correlation vs. explicit physics and the synthetic data bias mechanisms described above.

The most uncertain areas for a new project are: the optimal backbone architecture at the target scale (LLaMA3 vs. DiT vs. AR); the optimal compression ratio and design for the target use case; and the specific hyperparameter sensitivity landscape for the target model size. These uncertainties are addressable through targeted ablation studies, making them research opportunities rather than blockers.

6. Unresolved Questions and Decision-Critical Gaps

1. LLaMA3 vs. DiT vs. AR backbone superiority at target scale. Movie Gen's ablation claims LLaMA3 > DiT, but VideoAR demonstrates AR can match diffusion with 10× fewer inference steps. The debate is not settled at the architecture level — scaling effects may dominate either way. The specific evidence needed to resolve this is a multi-architecture controlled ablation across increasing compute scales (1e18 to 1e20 FLOPs), measuring whether the quality gap converges, which would indicate architectural equivalence, or remains constant, confirming a genuine advantage. Comparative studies at equal compute and data are needed to resolve this for any specific project.

2. Optimal video compression design at the target scale. The tradeoff between compression ratio (more compression = fewer tokens = faster but potentially lower quality) is domain-dependent. REDUCIO (64× latent reduction) and DC-VideoGen's (32×/64× spatial compression) demonstrate that much deeper compression than currently standard is viable, while ARVAE demonstrates that the way temporal and spatial information is decomposed (decoupled via optical flow) may matter as much as the raw compression ratio. The 2.5D vs. 3D tradeoff is a specific instance of this broader question.

3. Hyperparameter scaling for very large video models. Video DiT scaling laws (Paper 9) show that LR/BS sensitivity differs from language models, requiring per-scale tuning. However, the optimal hyperparameter schedule for models at the 30B scale (Movie Gen's scale) is not well-characterized. This directly affects training stability — one of the lecture's key constraints.

4. Long-horizon world consistency and physical accuracy. Genie 2 maintains consistency for 10–20 seconds (up to 1 minute). Longer-horizon video generation for extended agent training (minutes to hours of simulation) is not yet demonstrated. Error accumulation and world consistency degradation over long horizons remain fundamental challenges that constrain the world simulator use case. The synthetic data bias mechanism (distribution shift, compounding errors, artifact exploitation) compounds this limitation for embodied AI applications.

5. Chinese text rendering in video generation. Explicitly identified as an open challenge in both the lecture and the literature. None of the 15 opened papers provide a satisfactory solution. This is a specific capability gap that affects deployment in Chinese-language markets.

6. Optimal compute allocation between compression, backbone, and text conditioning. The literature provides scaling laws for the backbone component, but the optimal allocation of a fixed compute budget across all three components is not well-understood. This is a practical decision-critical question for any new project.

7. Recommended Next Steps

1. Conduct a targeted ablation comparing LLaMA3-based vs. DiT-based backbones at the target scale. The existing ablation from Movie Gen provides a reference point, but it was conducted at their specific scale with their specific data. A project-specific ablation is needed to determine the optimal backbone choice for the target application and compute budget. The ablation should measure both quality (VBench) and efficiency (inference latency, memory footprint).

2. Evaluate the deep compression VAE approach for the target use case. DC-VideoGen's chunk-causal design with 32×/64× spatial compression and REDUCIO's 64× image-conditioned approach represent fundamentally different efficiency-quality tradeoffs than the 2.5D/3D approaches discussed in the lecture. Running AE-Adapt-V adaptation experiments on the target pretrained model would clarify whether deep compression is beneficial for the specific application.

3. Incorporate FluxFlow-style temporal augmentation into the training pipeline. This is a low-cost, architecture-agnostic data-level addition that improves temporal quality across U-Net, DiT, and AR backbones. The specific perturbation parameters (frame shuffle window size, dropout rate, speed variation range) should be tuned for the specific data domain.

4. Design a comprehensive evaluation plan using VBench++ as the primary framework. Given the evidence that holistic metrics (FVD, FID) are inconsistent with human judgment, the evaluation plan should use VBench++ dimensions as the primary quality metrics, supplemented by domain-specific dimensions relevant to the target application. This avoids the trap of optimizing for aggregate metrics that mask specific weaknesses.

5. Evaluate OmniWeaving's IntelligentVBench benchmark framework as a potential evaluation protocol for the target application. IntelligentVBench extends standard video generation evaluation to reasoning-augmented generation, covering compositional reasoning, spatial-temporal causality, and multimodal integration. Even a preliminary evaluation against this framework — without full implementation — would clarify whether the target application requires reasoning capabilities and identify the competitive landscape for unified multimodal video generation systems.

6. Conduct a literature review specifically focused on Chinese text rendering in video generation, organized across three search dimensions: (a) font-specific generation methods — dedicated neural rendering

approaches for generating specific character glyphs and typography styles in video output; (b) OCR-guided video text rendering — pipeline methods that first generate text regions via OCR or layout prediction and then render character sequences with controlled font, size, and style; (c) cross-lingual text rendering transfer — leveraging advances in text-to-image rendering (e.g., AnyText, TextDiffuser) that handle multilingual text and investigating whether these techniques transfer to the video domain. A structured search across these three dimensions would clarify which approach is most promising for the specific application context.

8. Key Risks, Caveats, and Evidence Boundaries

1. World model physics accuracy is fundamentally limited to visual correlations. Both Genie 2 and Cosmos learn visual correlations that approximate physics, not explicit physical laws. Policies trained in generated world simulators may exploit visual artifacts rather than learning genuine physical understanding. Any application of video generation as world simulator infrastructure should include explicit physics validation, not just visual quality metrics.

2. Comparative claims against closed-source models are based on limited information. SORA, Kling, and other closed-source models have not disclosed training details, benchmark methodology, or evaluation protocols. All comparative claims (including Movie Gen's "beats SORA modestly") are based on limited public disclosure and should be treated as approximate directional signals rather than precise measurements.

3. Hyperparameter sensitivity means scaling is not straightforward. Video diffusion models are significantly more sensitive to learning rate and batch size than language models. The scaling laws from Papers 8 and 9 provide useful frameworks, but per-scale hyperparameter tuning is required. Projects that skip this tuning step risk training instability or suboptimal quality.

4. Evaluation framework maturity does not extend to all dimensions of interest. VBench and VBench++ are well-validated for standard text-to-video quality dimensions, but world model evaluation (physics accuracy, causal consistency, long-horizon dynamics), reasoning evaluation (compositional understanding, counterfactual generation), and domain-specific quality dimensions are not yet standardized. Projects targeting these dimensions must design custom evaluation protocols.

5. Training compute estimates exclude infrastructure development costs. Open-Sora 2.0's $200k figure covers training compute but not the infrastructure development, tooling, and engineering effort required to build the training system. Actual project costs may be significantly higher than the compute-only estimates.

6. Evidence-transfer risk from image to video domains. Several foundational papers (DiT, Flow Matching) focus on image generation, with video-specific considerations not addressed. The mechanisms transfer well in practice (AdaLN, OT Flow Matching), but video-specific adaptations (temporal coherence, variable-length processing, motion dynamics) may reveal failure modes not present in the image domain.

7. Synthetic data bias propagation risk. Both FluxFlow's temporal augmentation and world model training on synthetic video may introduce systematic biases through three compounding mechanisms: distribution shift between generated and real environments; compounding error propagation in closed-loop training where errors accumulate rather than average out; and artifact exploitation where policies learn visual shortcuts rather than true physical understanding. Projects using synthetic data should include diversity and accuracy verification protocols and should not rely on synthetic-only training for embodied agent applications.