Foresight: Can Video Prediction Ground Language Model Reasoning?

A Negative Result and Benchmark Proposal

Adrian Obleton
https://github.com/a1j9o94

January 2026

Abstract

We investigate whether vision-language models (VLMs) can benefit from generating pixel-level video predictions as part of their reasoning process. Our hypothesis: an AI that can "see" predicted outcomes through video generation will make better decisions than one reasoning purely in text or latent space. Through systematic experimentation across four research phases, we find that **(1)** VLMs cannot predict future states in their latent space, **(2)** video models can generate plausible continuations that VLMs understand, but **(3)** pixel-level verification (LPIPS) does not correlate with semantic correctness, preventing effective self-correction loops. We release our experimental framework as a benchmark for evaluating future video-language systems, proposing that the ability to predict and verify visual futures may be a key capability gap in current AI systems worth tracking over time.

1 Introduction

The remarkable progress in large language models (LLMs) has led to systems capable of sophisticated reasoning across many domains. However, these models reason primarily through text, lacking the ability to "imagine" and visualize outcomes before committing to predictions. Humans frequently engage in mental simulation—imagining what will happen if we take an action—as a core component of planning and decision-making.

This paper investigates a natural question: **Can AI systems benefit from generating pixel-level video predictions as part of their reasoning process?** We term this approach *Generative Latent Prediction* (GLP), combining the semantic understanding of vision-language models with the temporal modeling capabilities of video generation systems.

Our hypothesis was that a system which generates video predictions and compares them against actual outcomes could:

1. Detect when its predictions violate physical constraints (visible as artifacts)
2. Self-correct through verification loops
3. Achieve higher accuracy than pure text/latent reasoning

This approach differs fundamentally from latent-only prediction methods like V-JEPA [Bardes et al., 2024], which predict in learned representation spaces without generating interpretable pixels. Our bet was that pixel-level grounding would provide a richer error signal.

Summary of Results: After extensive experimentation, we find that current models are not capable of this form of grounded reasoning:

- VLMs cannot predict future states from their latent representations (7 architectural variations tested, all failed)

- Video models can generate plausible continuations, and VLMs can understand them (93% retention of action understanding)
- However, perceptual similarity metrics (LPIPS) do not correlate with semantic correctness ($r = 0.106$, AUROC=0.386)
- Verification loops achieve only 7.4% correction rate (below 15% threshold)

We release our experimental framework and propose it as a **benchmark** for tracking progress in video-language reasoning capabilities over time.

2 Related Work

2.1 Vision-Language Models

Recent VLMs like Qwen2-VL [Wang et al., 2024], LLaVA [Liu et al., 2023], and GPT-4V [OpenAI, 2023] achieve strong performance on visual understanding tasks. These models encode images into latent representations that are processed alongside text. However, their internal representations are optimized for understanding, not prediction—a distinction our experiments make explicit.

2.2 Video Generation Models

Diffusion-based video models including Stable Video Diffusion [Blattmann et al., 2023], LTX-Video [Lightricks, 2024], and Sora [Sora Team, 2024] can generate temporally coherent videos. These models excel at producing plausible visual continuations but are not designed to predict *what will happen* in a semantically meaningful way for downstream reasoning.

2.3 World Models and Latent Prediction

World models [Ha and Schmidhuber, 2018, Hafner et al., 2020] learn to predict future states in latent space for reinforcement learning. V-JEPA [Bardes et al., 2024] extends this to self-supervised learning, predicting masked video regions in latent space without pixel generation. Our work investigates whether pixel-level prediction provides advantages over latent-only approaches.

2.4 Self-Verification in Language Models

Self-consistency [Wang et al., 2023] improves LLM reasoning by sampling multiple reasoning paths and selecting the most consistent answer. Verification has proven valuable in code generation (execute and check) and mathematical reasoning (verify answer satisfies constraints). We investigate whether pixel-level verification can provide similar benefits for visual reasoning.

2.5 Perceptual Similarity Metrics

LPIPS [Zhang et al., 2018] measures perceptual similarity using deep network features, correlating well with human judgments. FVD [Unterthiner et al., 2019] extends this to video using I3D features. We investigate whether these metrics can serve as verification signals for prediction correctness.

3 Method

3.1 System Architecture

Our Foresight system combines three components:

1. **Hybrid Encoder:** DINOv2-ViT-L [Oquab et al., 2024] for spatial features + Qwen2.5-VL-7B for semantic understanding, fused via cross-attention
2. **Video Generator:** LTX-Video for image-to-video generation
3. **Verification Module:** LPIPS-based comparison between predicted and actual outcomes

3.2 Research Phases

We structured our investigation into four phases, each with explicit success criteria:

Phase 1: Reconstruction. Can VLM latents support video reconstruction? We found VLM latents preserve semantics but lose spatial precision (IoU=0.559 $<$ 0.6 threshold). This led to the hybrid encoder design combining DINOv2 spatial features with VLM semantics.

Phase 2: Bridging. Can we efficiently connect VLM to video decoder? A 10M parameter adapter achieved better quality than a 100M adapter (param_efficiency=1.165), validating efficient bridging.

Phase 3: Prediction. Can VLMs predict future states? Seven architectural variations (single-frame, multi-frame, temporal transformer, contrastive learning, pixel feedback with frozen/fine-tuned VLM) all failed to beat a copy baseline. This led to a "Video Predicts $\rightarrow$ VLM Describes" pivot.

Phase 4: Verification. Does pixel verification improve accuracy? This is the focus of our main results.

3.3 Datasets

We use Something-Something v2 [Goyal et al., 2017] (220,847 videos, 174 fine-grained action classes) as our primary evaluation dataset, as it requires understanding action semantics rather than just object recognition.

4 Experiments

4.1 Phase 3: VLM Cannot Predict Future States

Table 1: VLM prediction experiments. All architectures fail to beat the copy baseline.

Experiment	Architecture	cos_sim	Copy	Δ
E3.2	Single-frame	0.941	0.979	-0.038
E3.4	Multi-frame (8)	0.930	0.975	-0.045
E3.5	Temporal Transformer	0.930	0.975	-0.045
E3.6	Contrastive Loss	0.477	0.860	-0.384
E3.7a	Pixel Feedback (frozen)	0.209*	0.070*	-0.139
E3.7b	Pixel Feedback (LoRA)	$\sim$0.17*	$\sim$0.07*	negative

*L1 pixel loss (lower is better); all others cosine similarity (higher is better)

All architectures converge to approximately 0.93 cosine similarity regardless of temporal context or modeling approach. This strongly suggests VLM latent spaces do not encode future-predictive information.

Pivot: Video Predicts → VLM Describes. We pivoted to using video models for prediction and VLMs for understanding. LTX-Video generates future frames; VLM describes the generated content. Results:

- Temporal coherence ratio: 0.89 (vs 0.37 with simple extrapolation)
- VLM action recall on generated video: 70% (vs 75% on real video)
- Retention rate: 93%—VLM understands generated content almost as well as real video

4.2 Phase 4: Pixel Verification Does Not Work

With working video generation and VLM understanding, we tested whether pixel-level verification enables self-correction.

4.2.1 E4.1: Correlation Study

Question: Does LPIPS error predict whether predictions are semantically correct?

Table 2: LPIPS-Correctness correlation results.

Metric	Achieved	Target
Point-biserial correlation	0.106	> 0.30
p-value	0.58	< 0.05
AUROC	0.386	> 0.65
LPIPS gap (incorrect - correct)	-0.05	> 0.08

Key finding: LPIPS does not distinguish correct from incorrect predictions. The correlation is not significant ($p = 0.58$), and AUROC of 0.386 is *worse than random*. Surprisingly, incorrect predictions had *lower* LPIPS than correct ones.

4.2.2 E4.2: Calibration Study

Question: Does model uncertainty correlate with prediction error?

Table 3: Calibration study results.

Metric	Achieved	Target
Uncertainty-Error correlation	0.582	> 0.40
p-value	0.0007	< 0.05
Expected Calibration Error (ECE)	0.190	< 0.15
Reliability R^2	0.550	> 0.60

While uncertainty correlates significantly with error ($r = 0.582$, $p < 0.001$), the model is poorly calibrated (ECE=0.190 $>$ 0.15 threshold). Uncertainty does not reliably indicate when predictions are wrong.

4.2.3 E4.3: Verification Loop

Question: Can feedback enable self-correction?

Even the best feedback type (VLM description) achieves only 7.4% correction rate, well below the 15% threshold needed to justify the computational cost of verification loops.

Table 4: Verification loop results by feedback type.

Feedback Type	V1 Acc	V2 Acc	Correction Rate
Binary	10.0%	3.3%	0.0%
LPIPS score	10.0%	6.7%	3.7%
VLM description	10.0%	16.7%	**7.4%**
Target	-	-	$> 15\%$

5 Discussion

5.1 Why Did Pixel Verification Fail?

Our results suggest a fundamental disconnect between perceptual similarity and semantic correctness:

1. **Perceptually similar but semantically different:** Two videos can look similar at the pixel level while depicting different actions (e.g., pushing left vs right on a symmetric object).

2. **Perceptually different but semantically equivalent:** The same action can produce visually different results due to irrelevant variation (lighting, camera angle, object texture).

3. **LPIPS captures the wrong signal:** LPIPS is trained on human perceptual judgments, not task relevance. It may be sensitive to visual details that don't matter for semantic understanding.

5.2 What Would Need to Change?

For video-grounded reasoning to work, we would need:

1. **Task-specific verification metrics:** Train perceptual metrics on task-relevant differences rather than general visual similarity.

2. **Better video models:** Current models generate plausible but not necessarily accurate predictions. They need to actually predict what will happen, not just what could happen.

3. **Semantic verification:** VLM-based comparison (7.4% correction) outperformed LPIPS (3.7%), suggesting semantic rather than perceptual verification may be more promising.

5.3 Positive Findings

Despite the negative overall result, we made several useful discoveries:

- **Hybrid encoding works:** DINOv2 + VLM fusion achieves spatial IoU=0.837, solving the spatial information loss problem in VLM-only approaches.

- **VLMs understand generated video:** 93% retention of action understanding on LTX-Video outputs enables "Video Predicts $\rightarrow$ VLM Describes" workflows.

- **Small adapters suffice:** 10M parameter adapter outperforms 100M, suggesting efficient bridging is possible.

6 Benchmark Proposal: VideoReason

We propose releasing our experimental framework as a benchmark for tracking progress in video-language reasoning:

6.1 Tasks

1. **Future Prediction:** Given context frames and action description, predict the next N frames.
2. **Action Understanding:** Classify actions in both real and generated videos; measure retention rate.
3. **Verification Correlation:** Measure correlation between perceptual metrics and semantic correctness.
4. **Self-Correction:** Measure correction rate in verification loops with various feedback types.

6.2 Metrics

- **Temporal Coherence Ratio:** How realistic is generated motion?
- **VLM Retention Rate:** How well does VLM understand generated vs real video?
- **Verification AUROC:** Can perceptual metrics distinguish correct/incorrect?
- **Correction Rate:** Does verification enable self-improvement?

6.3 Why Track This Over Time?

Video generation models are improving rapidly. The capabilities we found lacking in 2026 may emerge in future systems. A standardized benchmark allows:

1. Tracking progress toward video-grounded reasoning
2. Comparing different video-language architectures
3. Identifying when the approach becomes viable

7 Conclusion

We investigated whether AI systems can benefit from generating pixel-level video predictions as part of their reasoning process. Our systematic experiments found that:

1. VLMs cannot predict future states from their latent representations
2. Video models can generate plausible continuations that VLMs understand
3. However, perceptual similarity does not correlate with semantic correctness
4. Verification loops do not enable effective self-correction

This negative result is valuable: it establishes that pixel-level verification, at least with current models and metrics, does not provide the error signal needed for grounded visual reasoning. The disconnect between perceptual and semantic similarity appears fundamental.

We release our framework as a benchmark for future video-language systems. The ability to predict and verify visual futures may be a key capability gap worth tracking as models improve. We encourage the community to revisit these experiments as new video generation and vision-language models emerge.

Acknowledgments

We thank the developers of Qwen2-VL, LTX-Video, DINOv2, and the Something-Something v2 dataset for making their work available.

References

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. (2024). V-JEPA: Latent video prediction for visual representation learning. *arXiv preprint arXiv:2402.04627*.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorber, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. (2023). Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. (2017). The "something something" video database for learning and evaluating visual common sense. In *ICCV*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *ICML*.

Ha, D. and Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. (2020). Dream to control: Learning behaviors by latent imagination. In *ICLR*.

Lightricks (2024). LTX-Video: Real-time video generation. `https://github.com/Lightricks/LTX-Video`.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. In *NeurIPS*.

OpenAI (2023). GPT-4V(ision) system card. Technical report.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2024). DINOv2: Learning robust visual features without supervision. *TMLR*.

Sora Team (2024). Video generation models as world simulators. Technical report, OpenAI.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. (2019). FVD: A new metric for video generation. *arXiv preprint*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. (2024). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

A Detailed Experiment Results

A.1 Phase 1: Reconstruction Quality

Table 5: Hybrid encoder ablation results.

Configuration	Spatial IoU	LPIPS	mAP@0.5	Latency
VLM only	0.559	0.264	0.001	baseline
DINOv2-ViT-B	0.742	0.198	0.095	+24%
DINOv2-ViT-L	**0.837**	**0.162**	0.182	+32%
DINOv2-ViT-G	0.851	0.155	0.201	+68%

A.2 Phase 2: Adapter Efficiency

Table 6: Adapter scaling comparison.

Adapter Size	LPIPS	Training Time	Param Efficiency
10M	0.289	0.32x	**1.165**
50M	0.312	0.65x	0.952
100M	0.346	1.00x	1.000

A.3 Infrastructure

All experiments were run on Modal cloud infrastructure using NVIDIA A100-80GB GPUs. Total compute: approximately 200 GPU-hours across all experiments.

B Benchmark Code

Code and evaluation scripts are available at: `https://github.com/a1j9o94/foresight`