Video-GPT

via Next Clip Diffusion

5-minute English report for ICLR 2026

Shaobin Zhuang et al.

Core message: treat each video clip as a visual word, then pretrain with next clip diffusion.

Self-supervised on 70M unlabeled videos
Strong forecasting + strong transfer

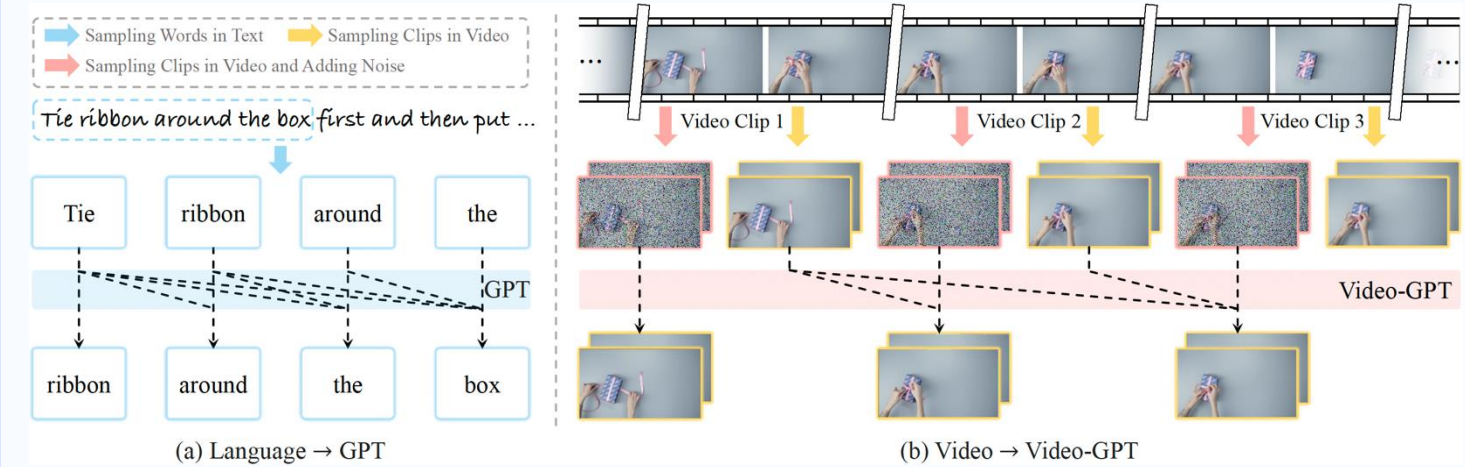


Figure 1: next clip diffusion analogy from the paper

AR across clips

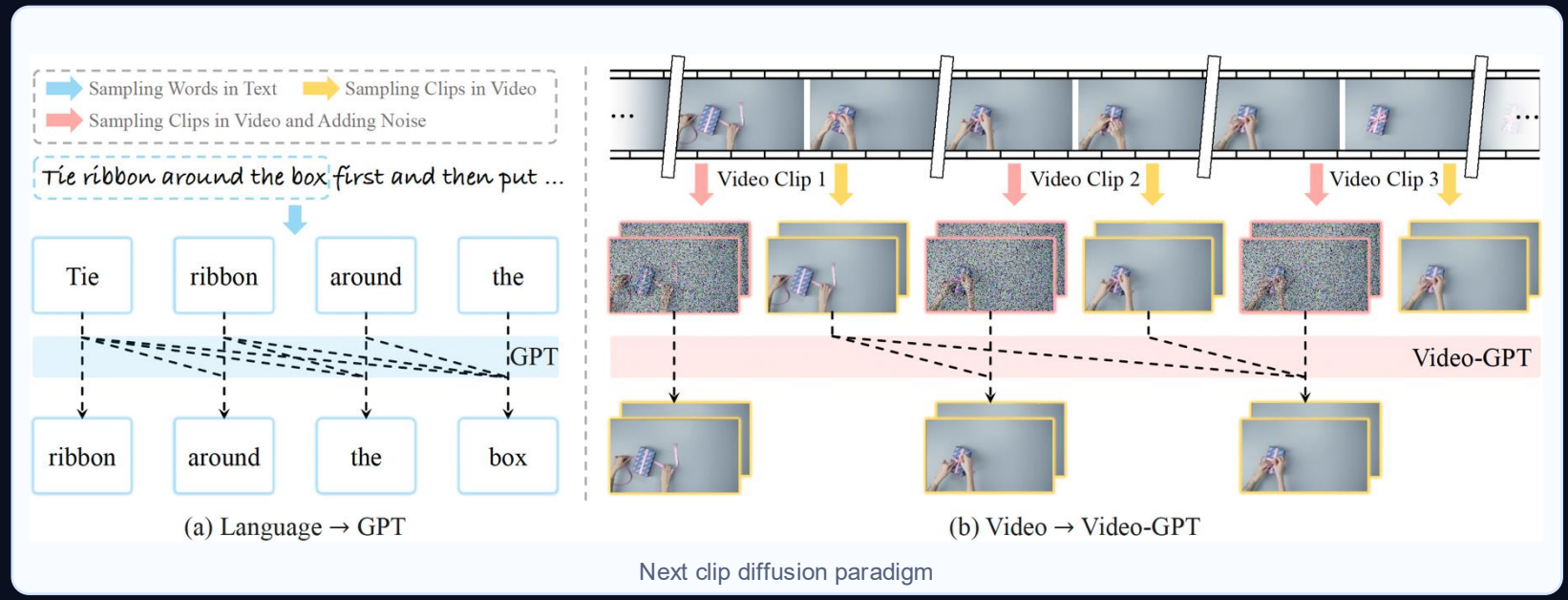
Diffusion within clip

Why Treat Video as a New Language?

From next-token prediction to next-clip prediction

- Language captures high-level abstraction, but misses rich spatial-temporal detail.
- Video naturally records physical dynamics, motion, and interaction over time.
- The paper treats each clip as a "visual word" and predicts the future clip from clean history.
- This combines GPT-style sequencing with diffusion-quality generation.

Key intuition: use clean past clips as the correct temporal context, then denoise the next noisy clip.



Next clip diffusion paradigm

Pretraining: Next Clip Diffusion

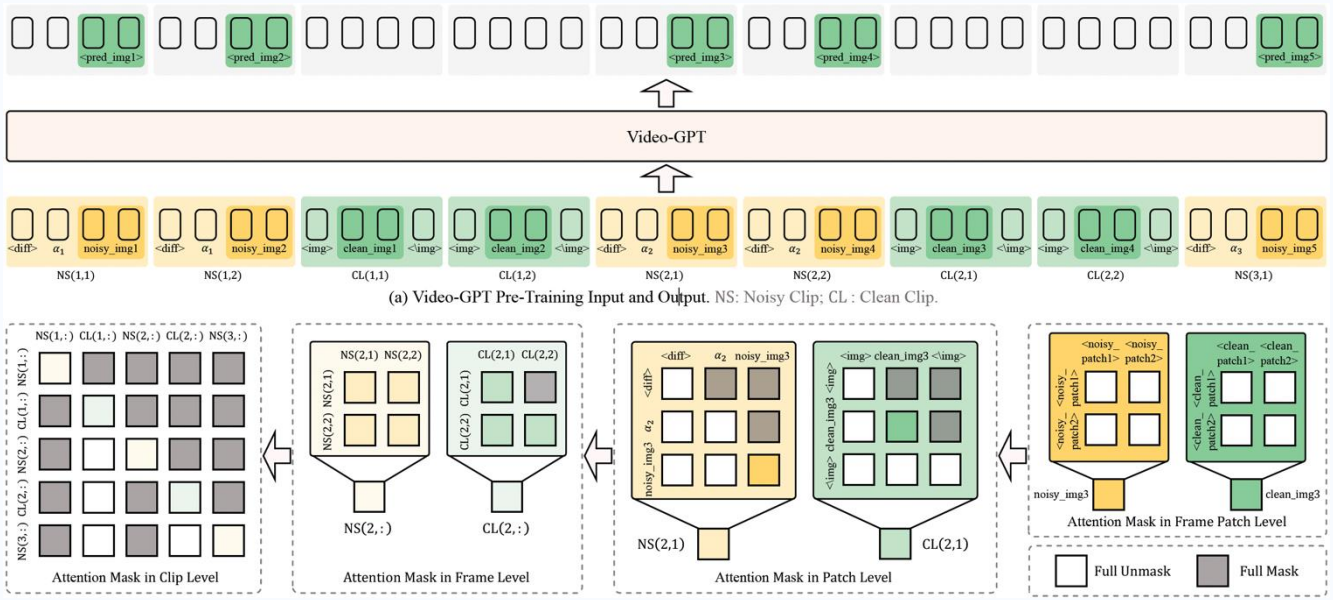
Interleaved noisy-clean clips + hierarchical attention masking

Training pipeline

- Sample video frames and divide them into K clips.
- Add diffusion noise in latent space to each clip.
- Build an input sequence: noisy clip, clean clip, noisy clip, clean clip, ...
- Predict the denoised next clip with previous clean clips as history.
- Train progressively from shorter videos to longer videos.

Hierarchical masking

Clip level: causal across clips
Frame level: causal for clean frames, bidirectional inside current noisy clip
Patch level: full spatial interaction within a frame



Video-GPT pretraining framework

Stage	Resolution	Frame Num.	Frame Interval	Clip Num.	Train Steps
1	Flexible	16	4	16	300000
2	Flexible	48	4	$\sim$ Uniform$\{2, 3, \cdots, 48\}$	25000
3	Flexible	48	$\sim$ Uniform$\{4, 5, \cdots, 12\}$	$\sim$ Uniform$\{2, 3, \cdots, 48\}$	40000
4	Flexible	80	$\sim$ Uniform$\{4, 5, \cdots, 12\}$	$\sim$ Uniform$\{2, 3, \cdots, 80\}$	20000

Progressive training strategy

Inference and Broad Transfer

One pretrained backbone supports video forecasting and six downstream vision and multimodal tasks

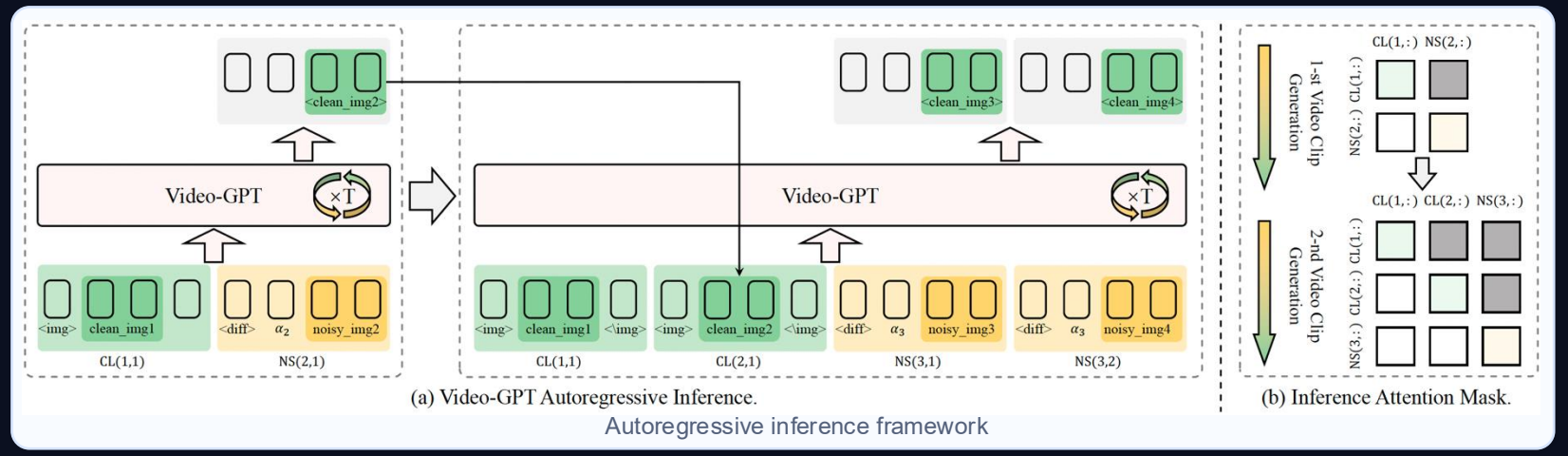
Inference logic

Start from an initial clean clip, denoise the next noisy clip, append it to history, and repeat with a sliding context window.

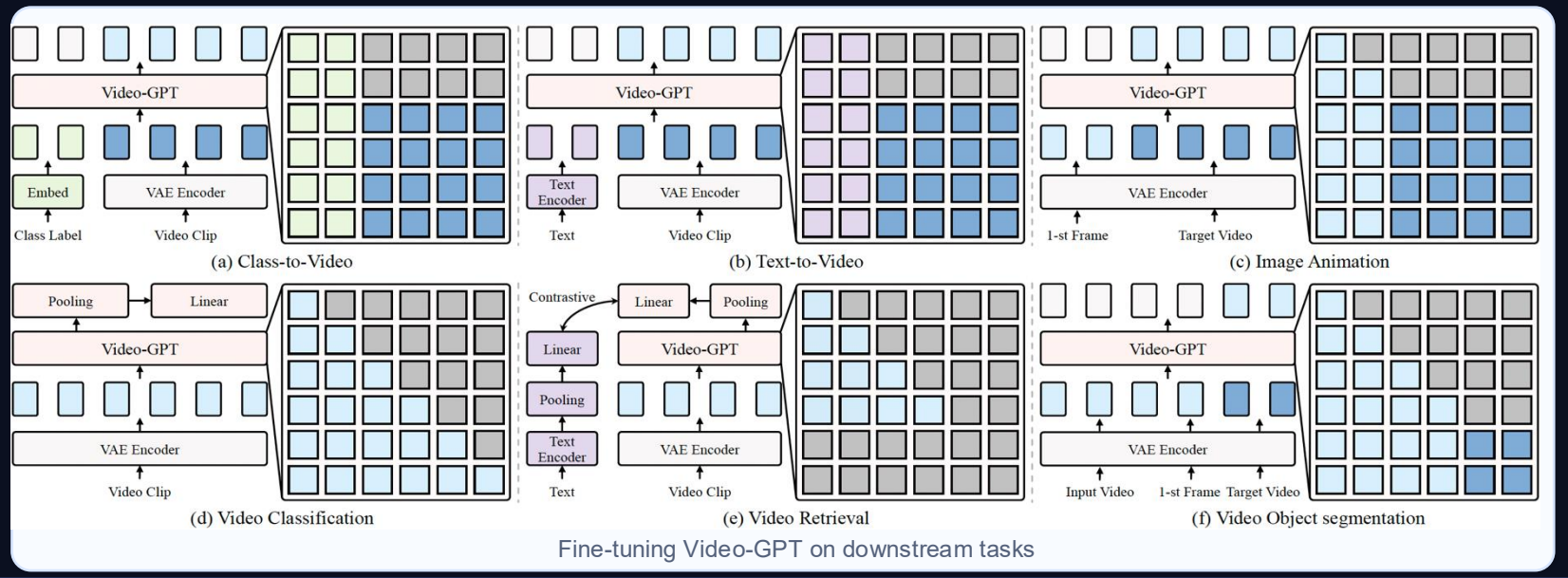
Downstream transfer

Generation: class-to-video, text-to-video, image animation
Understanding: classification, retrieval, object segmentation

This suggests reusable video representations, not just one narrow generation setup.



Autoregressive inference framework



Fine-tuning Video-GPT on downstream tasks

Main Results

State-of-the-art world modeling signal and strong downstream transfer

34.97
Physics-IQ score
Kling1.6: 23.64 | Wan2.1: 20.89

89.44
Kinetics-600 FVD(5000)
Best among compared methods in Table 3

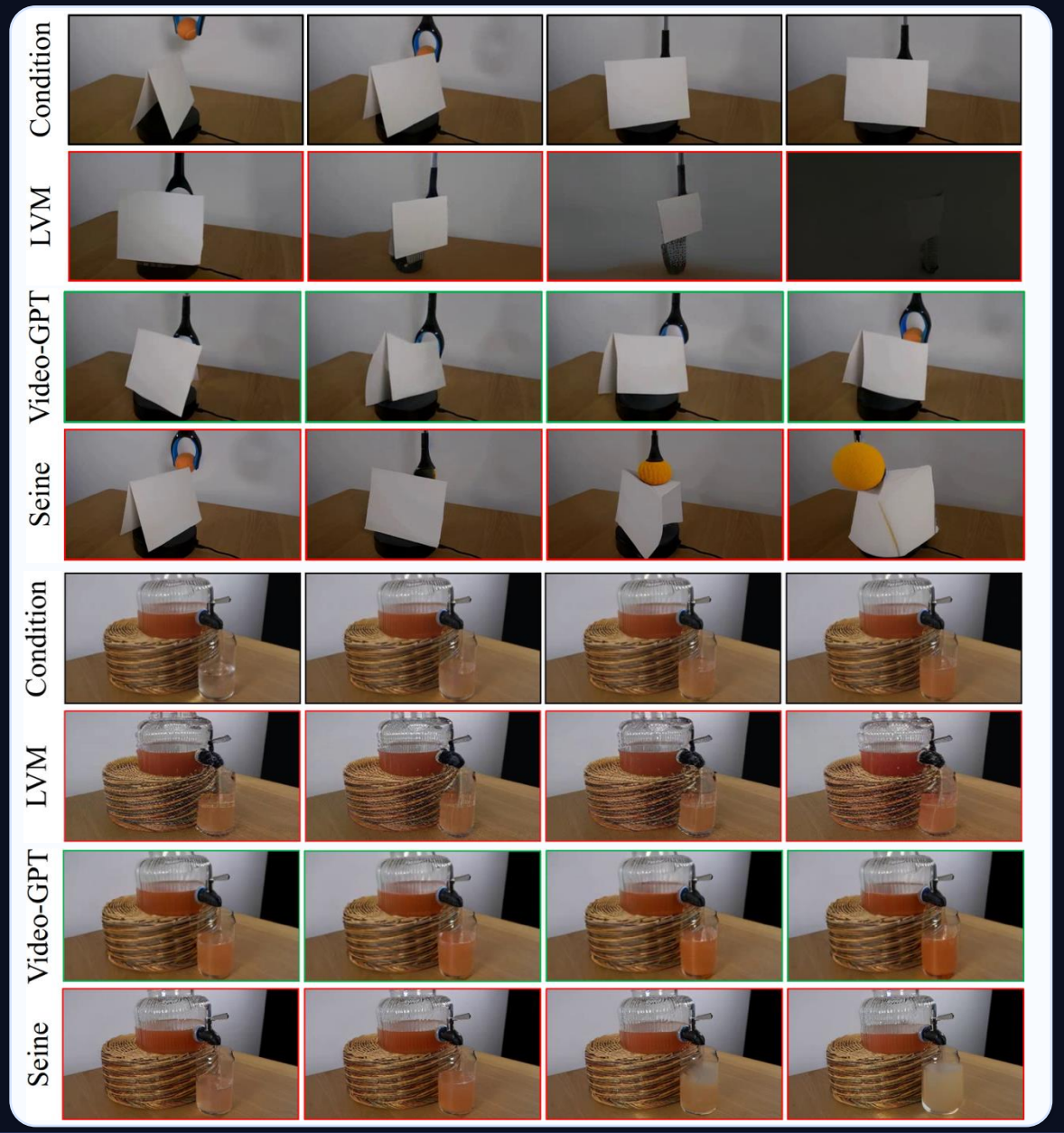
53
UCF-101 FVD
class-to-video

58.9
Top-1
video classification

22.8
R@1
MSR-VTT retrieval

What these numbers mean

- Strong deterministic forecasting on Physics-IQ
- Strong uncertain-video prediction on Kinetics-600
- Efficient transfer to both generation and understanding tasks



Takeaways

A concise foundation model that unifies GPT-style prediction and diffusion

- Clip = the basic predictive unit, analogous to a word in language.
- Diffusion improves local generation quality; autoregression preserves long-range temporal order.
- Large-scale self-supervised pretraining on Panda-70M enables broad transfer.
- Future work in the paper: multimodal pretraining and reinforcement-driven world interaction.

Main takeaway: next clip diffusion is a promising recipe for scalable video world models.