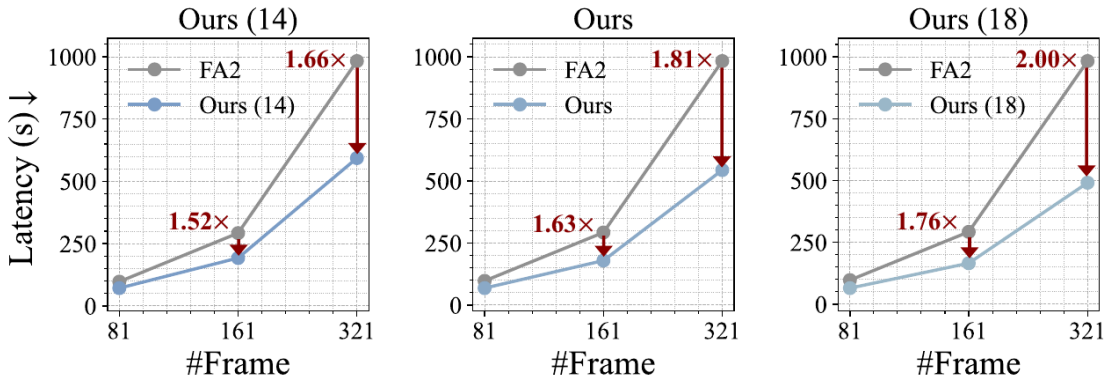
LinVideo: A Post-Training Framework towards O(n) Attention in Efficient Video Generation

Yushi Huang et al. • CVPR 2026

1. Why this problem matters

- Video diffusion models need very long sequences, so self-attention becomes the main inference bottleneck.
- The paper notes that a 10-second video can exceed 50K tokens.
- Sparse attention helps, but often still keeps more than half of dense-attention computation.
- Directly replacing all layers with linear attention usually hurts video quality.

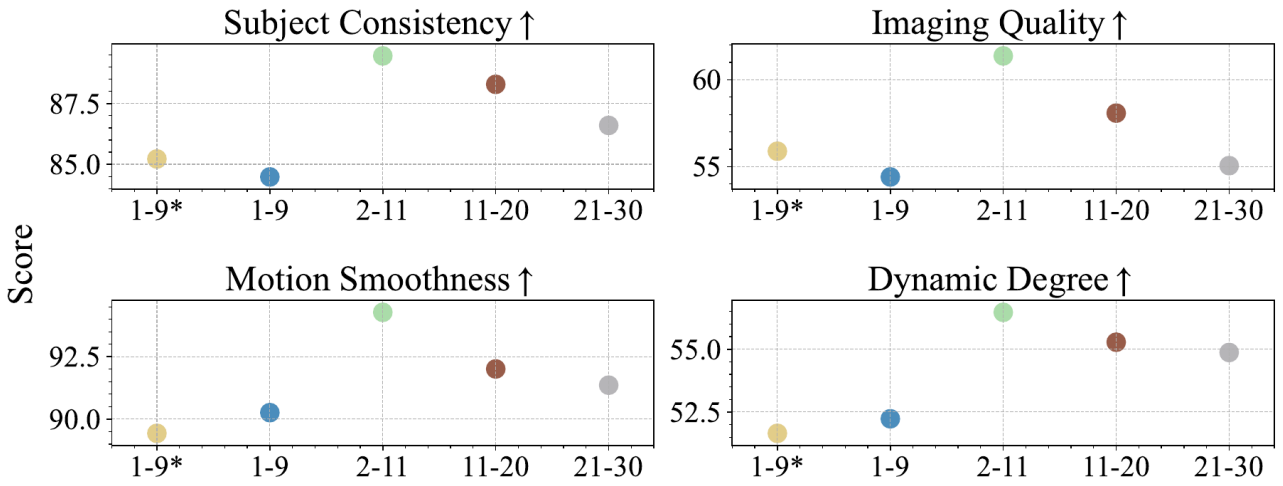


Question: can efficient post-training make linear attention practical for video generation?

Problem matters more as videos get longer.

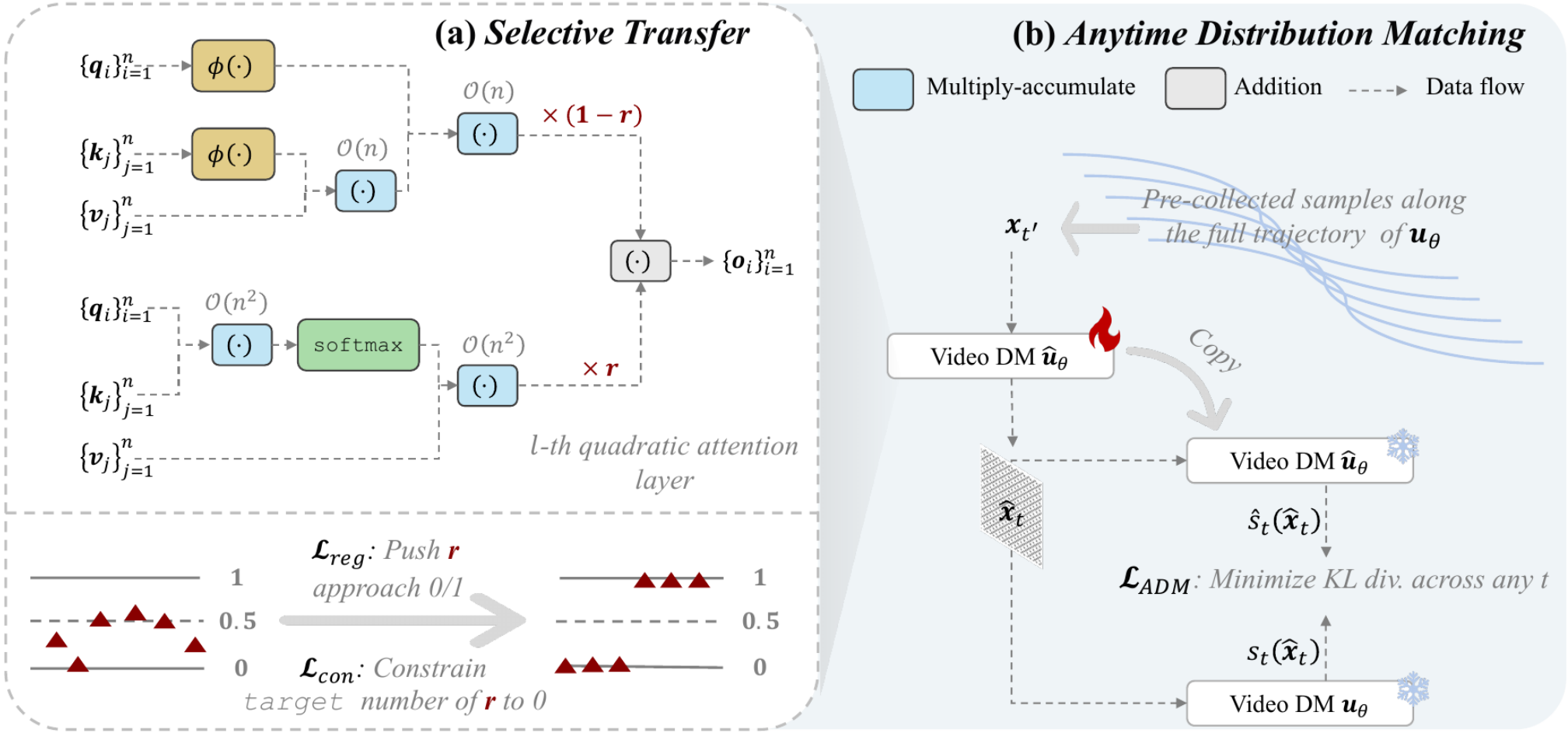
2. Observation: layer replaceability is uneven

- Some layer ranges can move to linear attention with little loss.
- Others damage subject consistency, image quality, and motion much more.



The problem is not only how to train linear attention, but also which layers to convert.

3. LinVideo overview

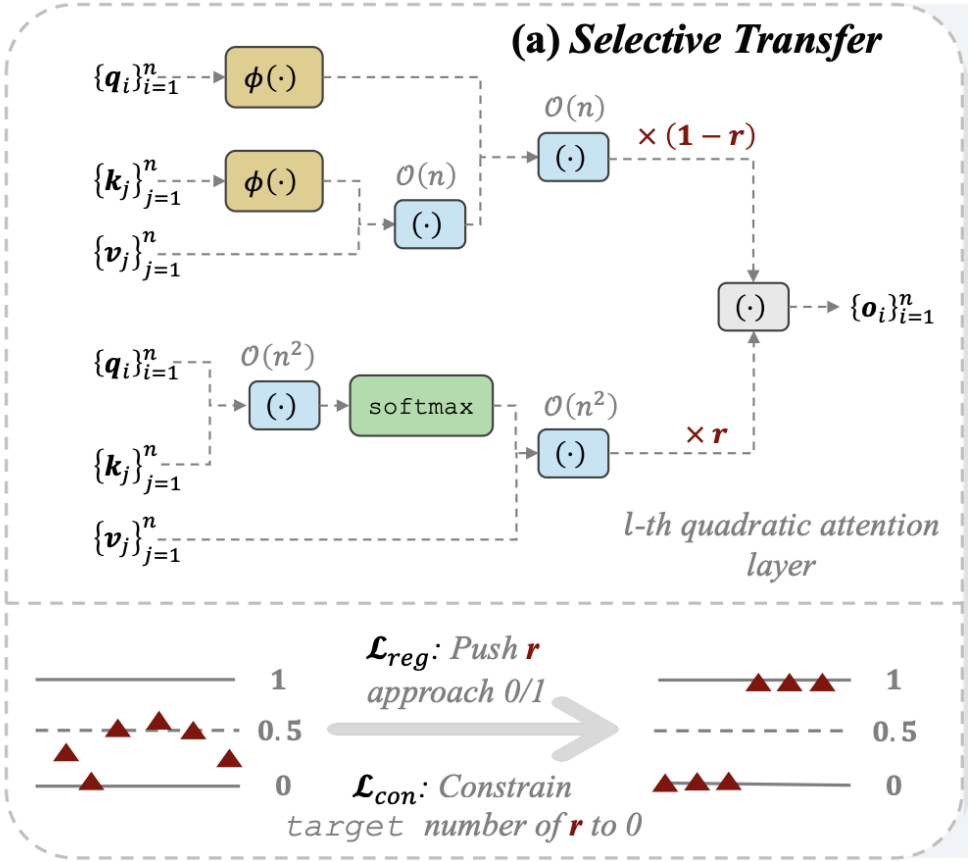


- Selective Transfer decides where to replace quadratic attention.
- ADM tells the converted model how to stay close to the original model during post-training.

4. Method I - Selective Transfer

- Learn a score r instead of hand-picking layers.
- Convert layers progressively, not all at once.
- Use a constraint + regularizer to end with near-binary decisions.

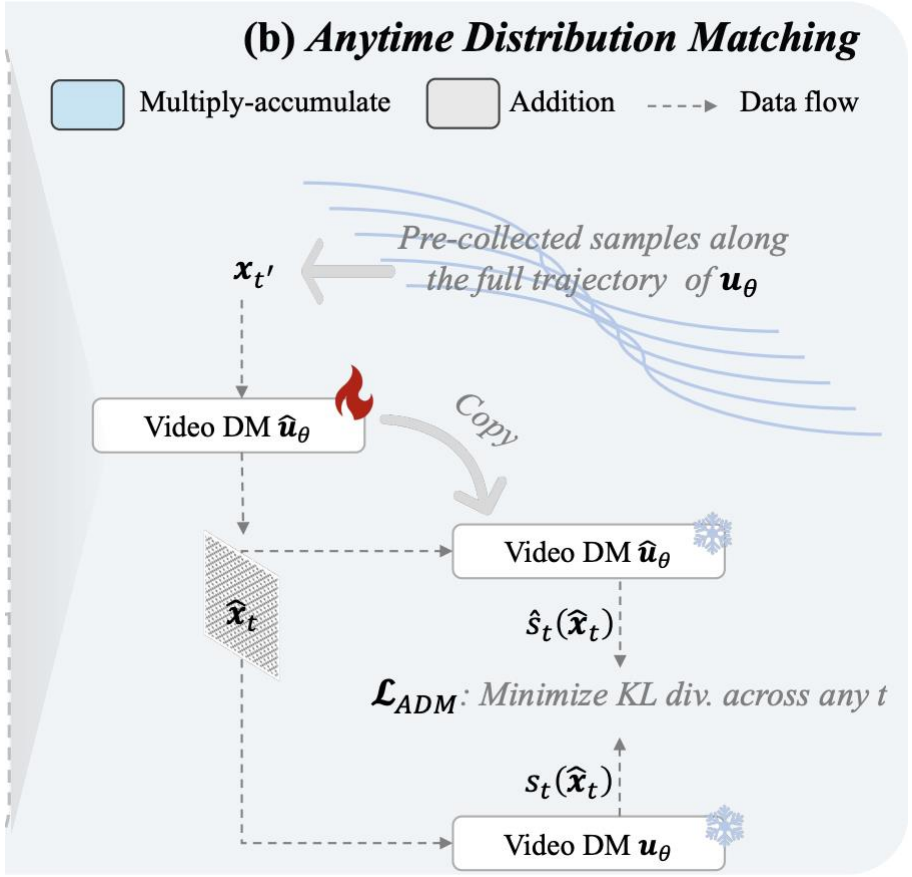
Layer selection becomes learnable, not heuristic.



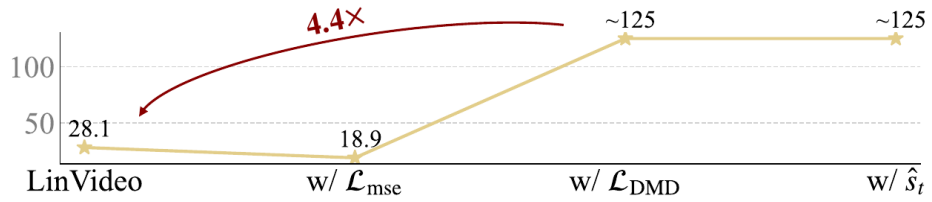
5. Method II - ADM

- MSE is too local.
- Few-step matching is expensive.
- ADM matches the original model along the whole denoising trajectory.

ADM keeps the converted model close to the original one along the full denoising path.



6. ADM is both better and faster



Method	Imaging Quality ↑	Aesthetic Quality ↑	Motion Smoothness ↑	Dynamic Degree ↑	Overall Consistency ↑
LinVideo	**66.07**	**59.41**	**98.19**	**59.67**	**26.52**
w/ $\mathcal{L}_{\text{mse}}$	61.56	56.37	96.32	52.48	21.46
w/ $\mathcal{L}_{\text{DMD}}$	57.44	52.79	90.72	49.37	16.96
w/ $\hat{s}_t^{\dagger}$	<u>65.61</u>	<u>59.34</u>	<u>97.82</u>	<u>59.43</u>	<u>25.87</u>

- The paper reports about 4.4× lower training time than DMD-style alternatives that need an extra score model.
- So ADM improves both quality and practicality.

7. Main quantitative results

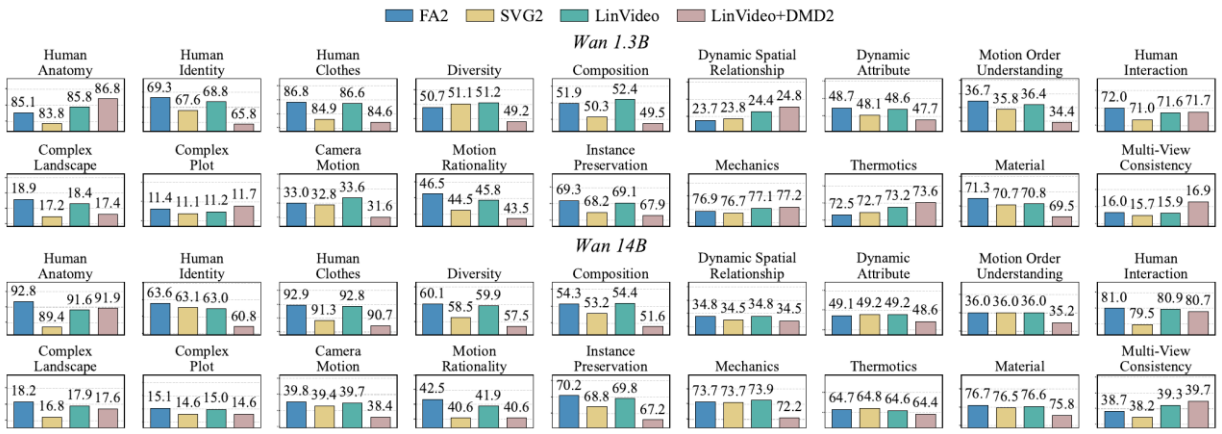
Wan 1.3B

- 1.43× speedup
- better benchmark quality than efficient baselines
- 15.9× with DMD2-distilled model

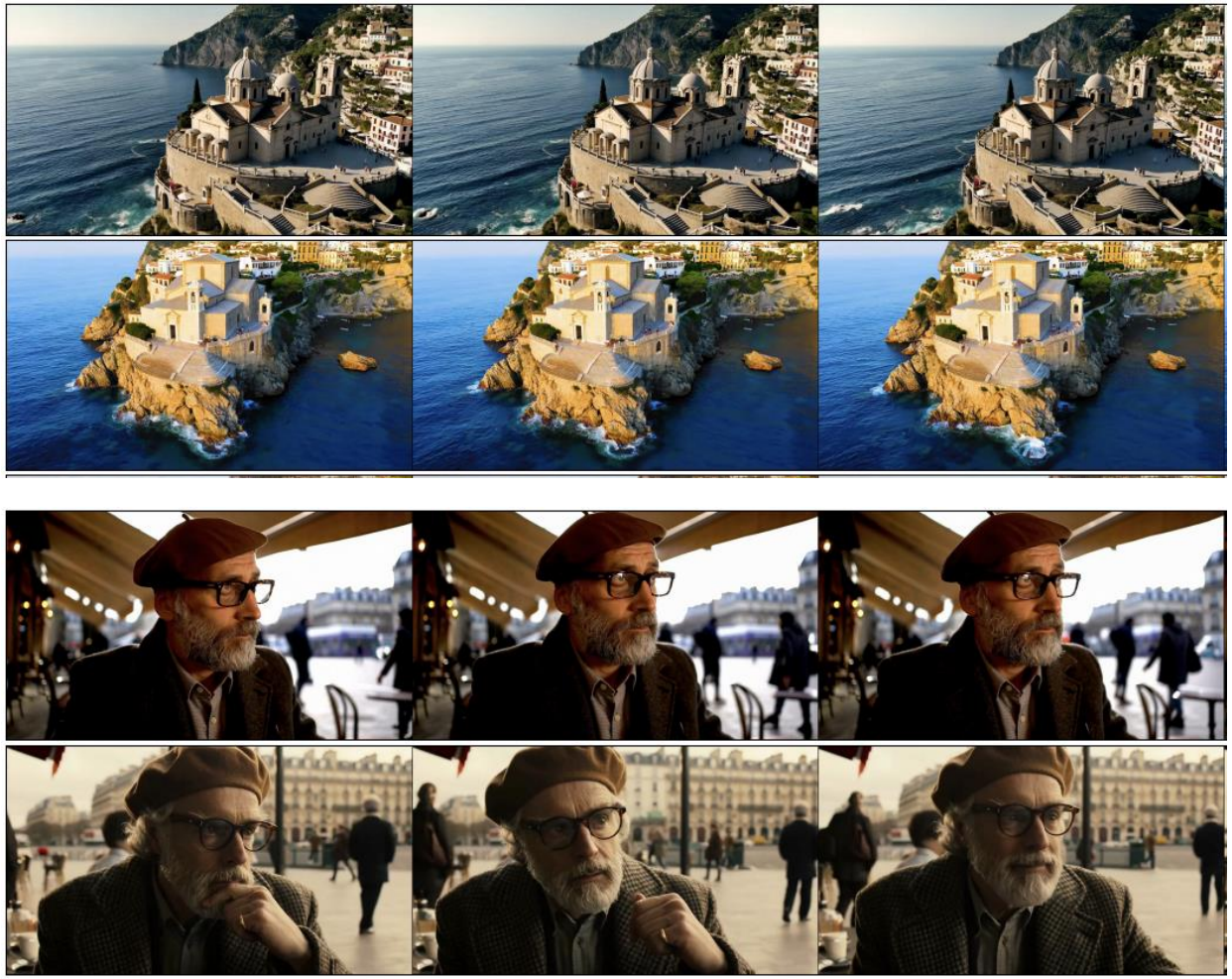
Wan 14B

- 1.71× speedup
- stronger quality than compared efficient methods
- 20.9× with DMD2-distilled model

Method	Latency (s)↓	Speedup↑	Imaging Quality↑	Aesthetic Quality↑	Motion Smoothness↑	Dynamic Degree↑	Background Consistency↑	Subject Consistency↑	Scene Consistency↑	Overall Consistency↑
Wan 1.3B (CFG = 5.0, 480p, fps = 16)										
FlashAttention2 [5]	97.32	1.00×	66.25	59.49	98.42	59.72	96.57	95.28	39.14	26.18
DFA [60]	88.95	1.09×	65.41	58.35	98.11	58.47	95.82	94.31	38.43	26.08
XAttn [53]	83.51	1.17×	65.32	58.51	97.42	59.02	95.43	93.65	38.14	26.22
SVG [50]	74.52	1.31×	65.78	59.16	97.32	58.87	95.79	93.94	38.54	25.87
SVG2 [54]	84.91	1.15×	66.03	59.31	98.07	59.44	96.61	94.95	39.14	26.48
LinVideo	68.26	1.43×	**66.07**	**59.41**	**98.19**	59.67	**96.72**	**95.12**	**39.18**	**26.52**
LinVideo + DMD2 [56]	**6.110**	**15.9×**	65.62	57.74	97.32	**61.26**	95.47	93.74	38.78	25.94
Wan 14B (CFG = 5.0, 720p, fps = 16)										
FlashAttention2 [5]	1931	1.00×	67.89	61.54	97.32	70.56	96.31	94.08	33.91	26.17
DFA [60]	1382	1.40×	65.93	60.13	96.87	69.34	95.37	93.26	33.14	26.12
XAttn [53]	1279	1.51×	65.47	60.36	96.28	69.25	95.24	92.97	33.22	26.14
SVG [50]	1203	1.61×	66.09	60.86	96.91	69.46	95.35	93.18	33.46	26.07
SVG2 [54]	1364	1.42×	66.25	61.08	97.12	69.43	95.51	93.39	33.52	26.14
LinVideo	1127	1.71×	**66.47**	**61.36**	**97.24**	**69.82**	**96.34**	**93.68**	**33.72**	**26.16**
LinVideo + DMD2 [56]	**92.56**	**20.9×**	65.74	59.68	96.32	69.74	95.38	92.88	33.18	26.09



9. Qualitative behavior



- Converted LinVideo still keeps scene layout, color, and motion plausible.
- The examples support the paper's claim that selective, trajectory-aware transfer is visually robust.

LinVideo preserves quality where direct linear replacement usually breaks.

10. Takeaways

- Convert part of video-DM attention from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$.
- Selective Transfer chooses where to convert.
- ADM tells the converted model how to stay close to the original one.

1.43× - 1.71×
standard-model speedup

15.9× - 20.9×
with DMD2-distilled models

One-line message:
linear attention becomes practical when transfer is selective and trajectory-aware.

Thank you!