Appendix

This document supplements the main paper as follows:

- Sec. A provides more implementation details;
- Sec. B details the effect of hyperparameter α of Eq. (10);
- Sec. C provides additional ablation studies for α;
- Sec. D gives the derivation of score difference (*i.e.*, Eq. (13));
- Sec. E presents examples of temporal jitter incurred by naive objective (*i.e.*, Eq. (7));
- Sec. F shows the results of directly distilling the original DM into a 4-step model with linear attention;
- Sec. G provides detailed setups for heuristic search in Tab. 3;
- Sec. H provides additional comparison for DMD2 distilled few-step models;
- Sec. I presents ablation studies for the kernel function of linear attention;
- Sec. J provides a comparison between SLA and our method;
- Sec. K shows the comparison results for CogVideoX [55];
- Sec. L presents visualization examples of our LINVIDEO.

A. More Implementation Details

In this section, we provide additional implementation details. Before training, we first collect inputs and outputs from `Wan` [49] across different sampling steps using prompts from `OpenVid` [34]. Each input/output contains 81 frames (21 frames in latent space), with an original resolution of $480p$ for `Wan` 1.3B and $720p$ for `Wan` 14B. During training, we use `Pytorch FSDP` [69] with a warm-up phase covering $\frac{1}{10}$ of the total training epochs, and set the global batch size to 48 for `Wan` 1.3B and 64 for `Wan` 14B. The learning rate follows a cosine annealing schedule, starting from 1×10^{-4}. For other hyperparameters, we set $\lambda = 0.01$, and α follows a linear decay from 20 to 2. For evaluation, we sample 5 videos for each unaugmented text prompt across the 8 dimensions in VBench [21]. For VBench-2.0 [70], we generate 3 videos per augmented text prompt across all dimensions, except for the Diversity dimension, where we generate 20 videos for each prompt.

B. Effect of α

As mentioned in the main text, we annealing decay the parameter α from large to small. This encourages r to move more adaptively at the initial phase to improve the training loss, but forces it to $0.0/1.0$ in the later phase. To understand this behavior, we provide a visualization in Fig. I. As shown in this figure, when α is large, the function remains nearly constant (close to 1.0) from $r^{(l)} = 0.5$ extending towards both boundaries, resulting in near-zero gradients in most regions. This flat landscape allows r to explore more freely during initial training. Conversely, when α is small, the gradient remains significant across a broader range, effectively pushing r towards the boundaries (0.0 or 1.0) in later training stages.

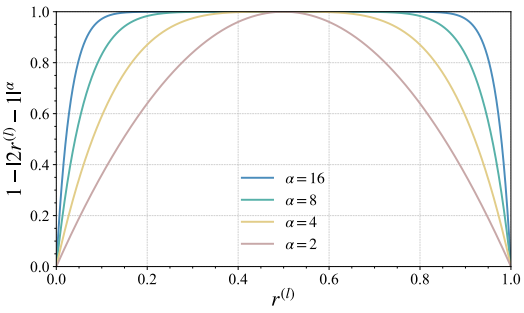


Figure I. Effect of α on $1-|2r^{(l)}-1|^{\alpha}$ of Eq. (10).

C. Ablation for α

Here, we provide additional ablation studies for α of $\mathcal{L}_{\text{reg}}$. As demonstrated in Tab. I, when we narrow or expand the range of α, the variation in all metrics stays within 0.03, which shows that our loss is not sensitive to changes in the chosen α values. In addition, this range of α yields strong final performance for models from 1.3B to 14B (Tab. 1), demonstrating good generalization. We also verify experimentally that using a dynamically changing α gives non-negligible improvements compared with using a fixed α, which confirms the benefit of our design: α encourages free exploration in the early stage of training and is gradually guided toward $0.0/1.0$ only in the later stage.

Table I. Ablation results of α. Our LINVIDEO employs $20 \rightarrow 2$ as the range of α.

Range of α	Imaging Quality ↑	Aesthetic Quality ↑	Motion Smoothness ↑	Dynamic Degree ↑	Overall Consistency ↑
$20 \rightarrow 2$	<u>66.07</u>	**59.41**	<u>98.19</u>	<u>59.67</u>	**26.57**
$16 \rightarrow 4$	**66.09**	59.38	98.16	**59.70**	26.54
$24 \rightarrow 1$	66.06	<u>59.40</u>	**98.20**	59.65	<u>26.56</u>
$4 \rightarrow 4$	65.87	59.08	98.10	59.08	26.27

D. Derivation of Eq. (13)

In this section, we provide a derivation of Eq. (13). Following Ma *et al.* [32] (their Eq. (9)), the score s_t associated with $\hat{\mathbf{x}}_t$ and estimated by $\boldsymbol{u}_\theta$ can be written as

$$s_t(\hat{\mathbf{x}}_t) = \sigma_t^{-1} \frac{\alpha_t \boldsymbol{u}_\theta(\hat{\mathbf{x}}_t) - \dot{\alpha}_t \hat{\mathbf{x}}_t}{\dot{\alpha}_t \sigma_t - \alpha_t \dot{\sigma}_t}, \tag{I}$$

where α_t, σ_t are the noise schedules and $\dot{\alpha}_t, \dot{\sigma}_t$ are their first-order derivatives with respect to t. For the *rectified flow models* [49] considered in this work, we take $\alpha_t = 1-t$ and $\sigma_t = t$. Substituting these schedules into the above expres-

sion yields

$$s_t(\hat{\mathbf{x}}_t) = \frac{1}{t}\frac{(1-t)\boldsymbol{u}_\theta(\hat{\mathbf{x}}_t) + \hat{\mathbf{x}}_t}{-t-(1-t)} = -\frac{1}{t}\big((1-t)\boldsymbol{u}_\theta(\hat{\mathbf{x}}_t) + \hat{\mathbf{x}}_t\big). \tag{II}$$

Likewise, the score $\hat{s}_t$ estimated by $\hat{\boldsymbol{u}}_\theta$ corresponding to $\hat{\mathbf{x}}_t$ is given by

$$\hat{s}_t(\hat{\mathbf{x}}_t) = -\frac{1}{t}\big((1-t)\hat{\boldsymbol{u}}_\theta(\hat{\mathbf{x}}_t) + \hat{\mathbf{x}}_t\big). \tag{III}$$

Therefore, the score difference is

$$s_t(\hat{\mathbf{x}}_t) - \hat{s}_t(\hat{\mathbf{x}}_t) = -\frac{1-t}{t}\left(\boldsymbol{u}_\theta(\hat{\mathbf{x}}_t) - \hat{\boldsymbol{u}}_\theta(\hat{\mathbf{x}}_t)\right). \tag{IV}$$

E. Temporal Jitter of Naive Objective

We provide examples in Fig. II using the naive training objective $\mathcal{L}_{\text{mse}}$ (*i.e.*, Eq. (7)), where the results clearly exhibit temporal jitter.

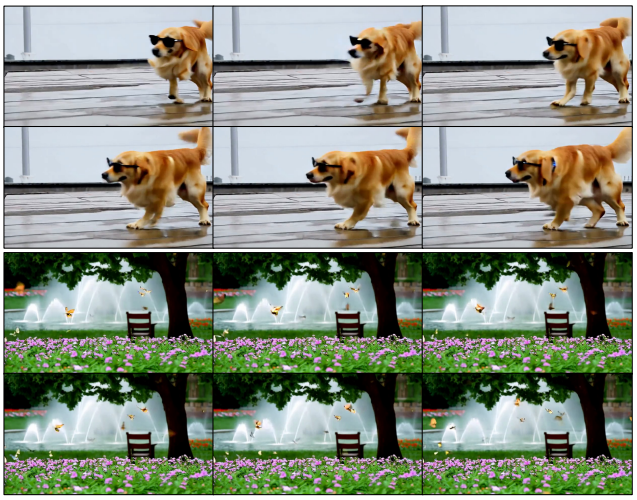


Figure II. A segment of 6 adjacent frames from an 81-frame, 480p, 16 fps video generated by the linear-attention `Wan 1.3B` [49] trained with the naive objective. Other settings are the same as those for LINVIDEO. Both the dog's legs (*Upper*) and the butterflies (*Lower*) exhibit severe temporal jitter.

F. Directly combine Few-step Distillation with Linear Attention

As shown in Tab. II, we observe that directly combining a few-step distillation with linear attention replacement (*i.e.*, our *selective transfer*) leads to severe performance degradation. This motivates a two-stage pipeline, *viz.*, LINVIDEO→DMD2 [56]. In the first stage, linear attention is introduced and already attains strong performance; in the second stage, few-step distillation is applied to further accelerate the model. This two-stage design stabilizes training and avoids the collapse observed with the direct combination.

Table II. Performance for the few-step distilled linear attention `Wan 1.3B` [49]. "LINVIDEO + DMD2" denotes that we first employ LINVIDEO to obatin a linear attention DM and then use DMD2 [56] to distill is to the 4-step version. "*ST* + DMD2" implies that we combine our *selective transfer* and DMD2 to obtain the 4-step linear attention model in a single training stage.

Method	Imaging Quality↑	Aesthetic Quality↑	Motion Smoothness↑	Dynamic Degree↑	Overall Consistency↑
LINVIDEO + DMD2	**65.62**	**57.74**	**97.32**	**61.26**	**25.94**
ST + DMD2	48.72	39.64	73.85	41.97	11.06

Table III. Comparison on `Wan 1.3B` (all methods use 4-step DMD2). Due to resource limits, we will add more results in the future.

Metric	FA2	DFA	XAttn	SVG	SVG2	LINVIDEO
VBench↑	67.68	64.71	64.32	65.83	66.06	**66.98**
Latency (s)↓	8.76	8.26	7.82	6.95	7.96	**6.11**

Table IV. Training FLOPs ($3K$ steps).

Method	FLOPs↓
LINVIDEO	5.97×10^4P
w/ $\mathcal{L}_{\text{mse}}$	$\mathbf{4.77\times10^4}$**P**
w/ $\mathcal{L}_{\text{DMD}}$	2.97×10^5P
w/ $\hat{s}_t$	2.97×10^5P

G. Details of Heuristic Search

For the *Heuristic* search method in the ablation study, we use 128 prompts from `OpenVid` [34]. Starting from the original DM, at each step we consider all remaining quadratic attention layers, construct modified variants where only one candidate layer is replaced by a linear attention layer, and measure the resulting output differences. We then permanently replace the layer that yields the smallest difference and repeat this procedure based on the updated model until the desired number of `target` layers has been selected.

H. Comparison for Few-step Models

Tab. III [9] shows that, when combined with DMD2, our method outperforms all baselines.

Table V. Comparison on CogVideoX-2B.

Metric	FA2	DFA	XAttn	SVG	SVG2	LINVIDEO
VBench↑	65.97	65.36	65.08	65.54	65.78	**65.97**
Latency (s)↓	41.35	37.84	34.29	33.13	38.46	**29.64**

Table VI. Comparison with SLA. "*" denotes results on an RTX5090 GPU; other results are on an H100 GPU.

Method	VBench↑	Latency (s)↓	Latency (s)↓*
`Wan 1.3B`	67.63	97.32	162.65
SLA	65.72	93.95	**78.42**
LINVIDEO	**67.61**	**68.26**	112.14

Table VII. Comparison across kernel functions on `Wan 1.3B`.

Kernel	VBench↑	Latency (s)↓
Hedgehog	**67.61**	68.26
`ReLU`	65.48	**68.12**
Taylor Exp	67.24	68.54

I. Ablation for Kernel Function

Hedgehog [63] preserves softmax-like *spiky weights* and *dot-product monotonicity*, making it more expressive than

[9] We report the average score of the chosen VBench dimensions Tab. 1. "FA2" denotes lossless FlashAttention2 baseline.

typical linear attention kernels. This benefits video generation (verify in Tab. VII [10]), where long sequences require sharp, stable long-range interactions.

J. Comparison between LINVIDEO and SLA

As shown in Tab. VI, LINVIDEO outperforms SLA on VBench. SLA's large speedup mainly comes from its hardware-specific kernels (currently only supporting RTX5090), while LINVIDEO uses `torch` implementation. Moreover, LINVIDEO can be further accelerated if we combine sparse attention as SLA (95% sparsity).

K. Comparison for CogVideoX

Our methods do not rely on any specific model architecture design. As shown in Tab. V, LINVIDEO outperforms baselines on CogVideoX.

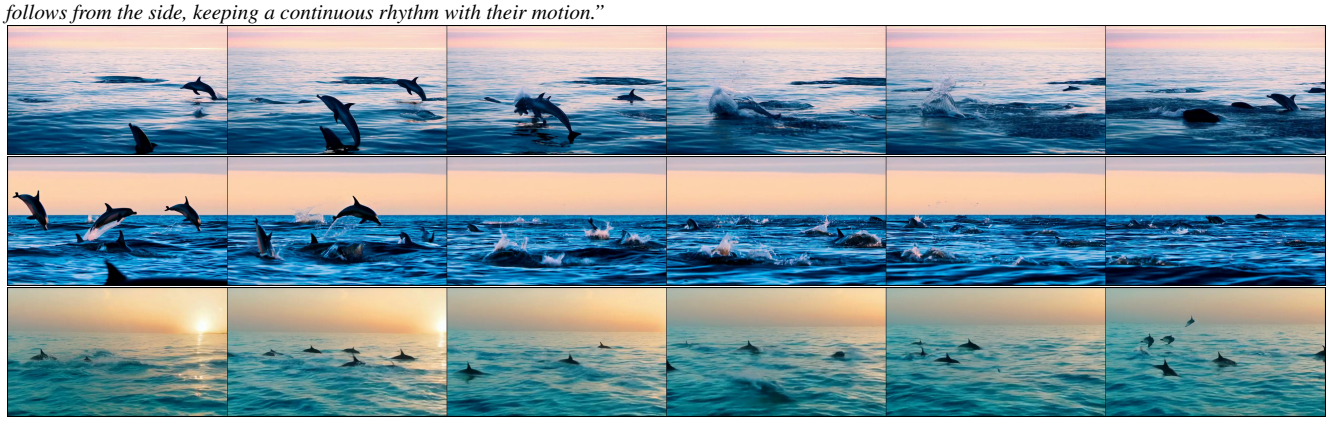
L. Visualization Results

In this section, we present randomly selected samples generated by LINVIDEO without any cherry-picking, as shown in Figs. III–IV. For a more detailed inspection, we recommend zooming in to closely examine the individual frames.

[10] Taylor Exp: refer to Zhang *et al.* [63].

Prompt: *"A volcano erupts in the distance, glowing lava rivers flowing against a darkened sky."*

Prompt: *"A pod of dolphins leaps out of the sparkling ocean in graceful arcs, splashing back into the water as the horizon glows with sunset; the camera follows from the side, keeping a continuous rhythm with their motion."*



Prompt: *"A massive elephant walks slowly across a sunlit savannah, dust rising around its feet, the warm glow of sunset illuminating the horizon; the camera moves steadily forward alongside, emphasizing the grandeur of its stride."*

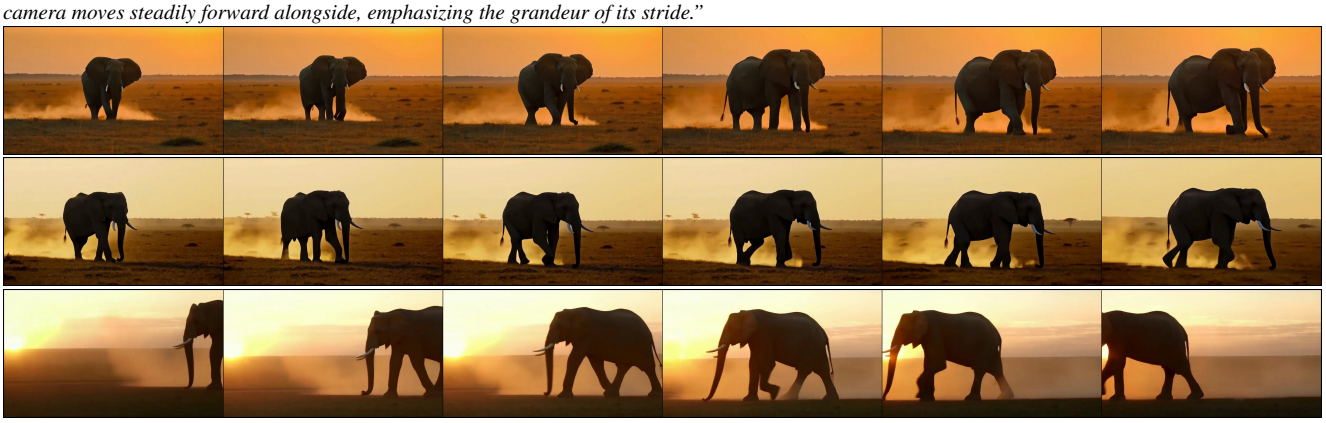
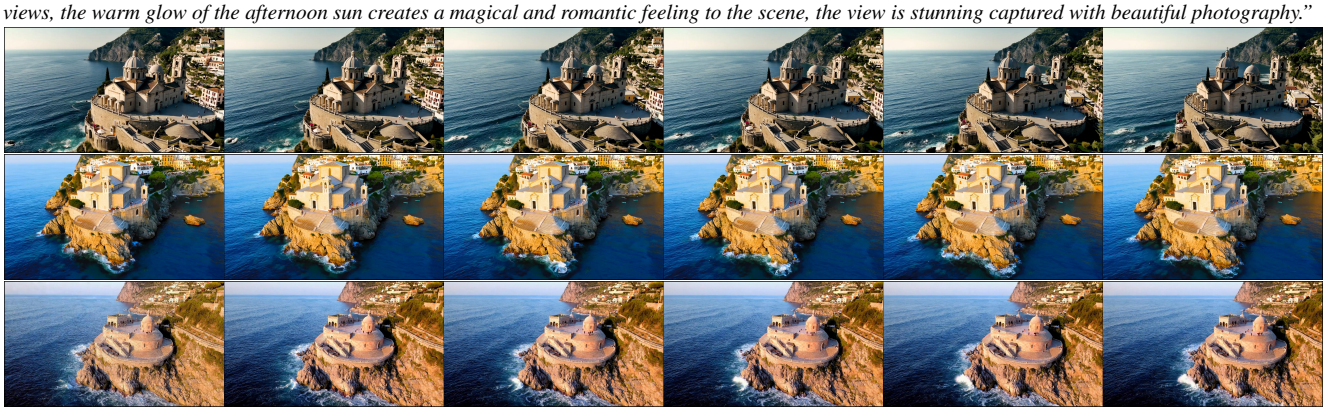


Figure III. Visual results at 480p across `Wan` 1.3B [49] (*Upper*), 1.3B LINVIDEO (*Middle*), and 1.3B LINVIDEO+ 4-step DMD2 [56] (*Lower*).

Prompt: *"Retro 80s Monster Horror Comedy Movie Scene: Color film, children's bedroom bathed in soft, warm light. Plush monsters of various sizes and colors are having a chaotic party, jumping on the bed, dancing to upbeat music, and throwing confetti. The walls are adorned with posters of classic 80s movies, and the room is filled with the playful laughter of children."*

Prompt: *"A drone camera circles around a beautiful historic church built on a rocky outcropping along the Amalfi Coast, the view showcases historic and magnificent architectural details and tiered pathways and patios, waves are seen crashing against the rocks below as the view overlooks the horizon of the coastal waters and hilly landscapes of the Amalfi Coast Italy, several distant people are seen walking and enjoying vistas on patios of the dramatic ocean views, the warm glow of the afternoon sun creates a magical and romantic feeling to the scene, the view is stunning captured with beautiful photography."*



Prompt: *"An extreme close-up of an gray-haired man with a beard in his 60s, he is deep in thought pondering the history of the universe as he sits at a cafe in Paris, his eyes focus on people offscreen as they walk as he sits mostly motionless, he is dressed in a wool coat suit coat with a button-down shirt, he wears a brown beret and glasses and has a very professorial appearance, and the end he offers a subtle closed-mouth smile as if he found the answer to the mystery of life, the lighting is very cinematic with the golden light and the Parisian streets and city in the background, depth of field, cinematic 35mm film."*

Figure IV. Visual results at 720p across `Wan` 14B [49] (*Upper*), 14B LINVIDEO (*Middle*), and 14B LINVIDEO+ 4-step DMD2 [56] (*Lower*).