# SEAE: SPECTRAL ENERGY AND ABLATION ERROR FOR STRUCTURAL PRUNING OF FEW-STEP VIDEO DIFFUSION TRANSFORMERS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The increasing size of Video Diffusion Models (VDMs) makes inference computationally expensive even after step distillation. Few-step distillation accelerates VDMs by reducing the denoising steps, but each step still executes the full width attention modules and the full depth Diffusion Transformer (DiT), leaving substantial memory usage and computation costs. Few-step VDMs Structural pruning is challenging because each solver step performs a larger portion of the noise-to-video transformation, giving pruning errors fewer subsequent steps in which to be corrected. We introduce SEAE, a structural pruning framework that reduces both attention width and DiT depth in few-step VDMs. Specifically, (**i**) our MHA-CSS method constructs Gram matrices from projected attention-head activations at all solver timesteps, uses their cumulative spectral energy to determine how many heads to retain in each self- and cross-attention module, and then selects the retained heads through CSS-inspired backward elimination with scalar refitting; (**ii**) we score each DiT blocks using Ablation Error, which temporarily removes each DiT block at a given timestep, completes the remaining sampling process, and measures its influence by comparing the final generated latent with the unablated reference. We form a low error candidate set at each solver timestep and retain their intersection as one static block pruning mask. Both masks are derived solely from forward passes over calibration set and define a fixed compact architecture shared across all timesteps, without runtime routing or cross-timestep feature caching. Recovery distillation finally trains the compact student using a weighted combination of teacher-student distillation and Flow Matching losses. Our model reduces the parameter count by 25.4% relative to the four-step Wan2.2-TI2V-5B teacher model, while nearly preserving its VBench-I2V total score for video generation.

## 1 INTRODUCTION

Video Diffusion Models (VDMs) generate temporally coherent videos from text and, in the text–image-to-video (TI2V) setting, an optional reference image. Recent systems build this capability on diffusion transformers (DiTs) (Peebles & Xie, 2023; Wan et al., 2025), whose spatio-temporal token sequences and billions of parameter backbones make every denoising evaluation expensive. Few-step distillation reduces the number of evaluations, but each remaining evaluation still executes the full denoiser. The resulting bottleneck is therefore not only the number of sampling steps, but also the amount of computation performed inside each of those few steps.

Currently, several acceleration strategies address parts of this bottleneck, mainly including structural pruning, feature cache, and dynamic pruning methods. Structural pruning permanently removes attention heads, FFN channels, and DiT blocks to obtain a model with fewer parameters, while video-specific methods introduce temporal or deployment-aware objectives (Fang et al., 2023; 2025; Ma et al., 2026a; Chavhan et al., 2026; Wu et al., 2025a; Karnewar et al., 2026; Wang et al., 2026; Ghafoorian et al., 2026). Feature caching methods such as MagCache (Ma et al., 2026b) and EasyCache (Zhou et al., 2025) instead reuse selected intermediate computations without changing the denoiser parameters. Similar to feature caching, dynamic pruning shortens inference time with the complete model parameters retained by skipping DiT blocks and merging similar tokens (Zhu et al.,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

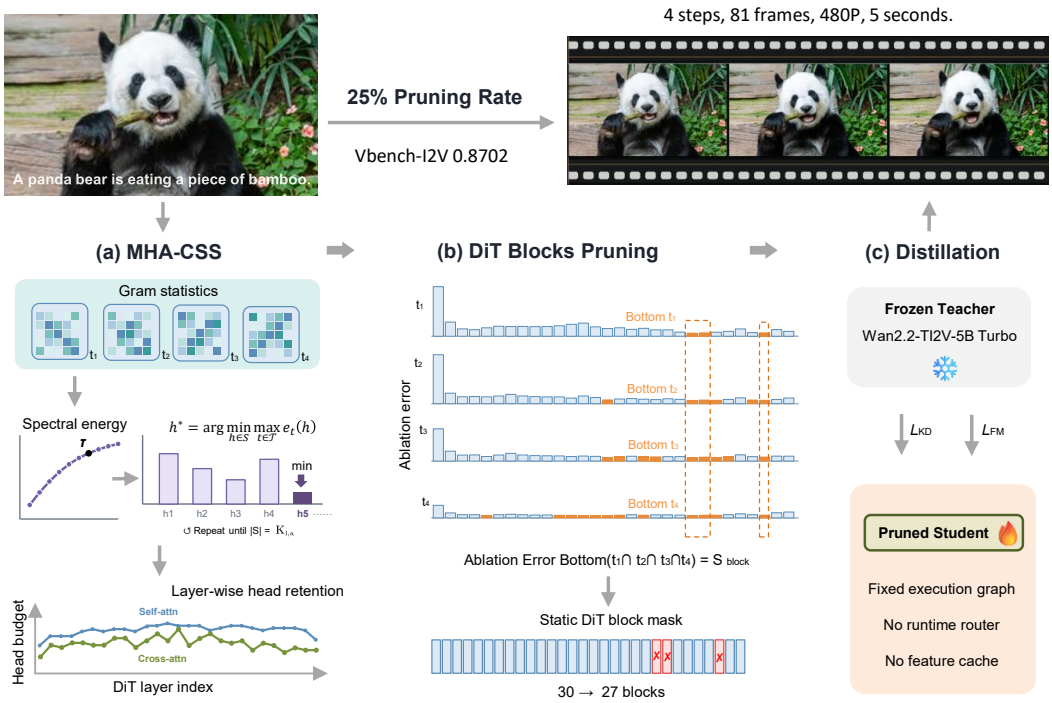


Figure 1: Overview of our TI2V pruning framework. We first apply MHA-CSS to determine module-specific head budgets and select retained heads, then use trajectory-level Ablation Error to identify timestep-consistent DiT blocks for static pruning. Recovery distillation training combines teacher-student knowledge distillation and a flow-matching objective.

2024; Zhang et al., 2025). These approaches serve different deployment goals: feature caching and dynamic pruning reduce executed latency along a sampling trajectory, whereas structural pruning can produce a model with both fewer parameters and lower latency.

However, few-step models introduce a distinct regime for structural pruning. Since only a few denoising steps are available, each step must carry a larger fraction of the transformation from noise to video latent, leaving less opportunity for perturbation-induced errors to be corrected along the trajectory. Structural pruning is therefore attractive for further reducing latency, but it can also perturb both per-frame appearance and cross-frame motion in a highly coupled manner. Existing evidence is consistent with this concern, although it does not yet provide a controlled multi-step/few- step comparison. VDMini shows that a one-step video model cannot be pruned uniformly: shallow and deep layers contribute differently to frame content and motion dynamics, motivating depth-aware pruning and motion consistency recovery (Wu et al., 2025a). In a broader vision generative setting, EcoDiff explicitly notes that step-distilled models are harder to fine-tune because their mappings are non-smooth (Zhang et al., 2026). By contrast, conventional multi-step diffusion pruning has reported approximately 50% FLOPs reduction and pruning rate (Fang et al., 2023; 2025). These observations motivate a pruning method that analyzes few-step model redundancy at a higher granularity, such as at the attention-module and timestep levels, to obtain a more compact model.

Our experiments reveal a further structural asymmetry in the studied VDMs: attention modules and DiT blocks are substantially more removable and less sensitive than FFN modules, consistent with observations from existing efficient VDMs research (Wu et al., 2025b). We therefore prune the MHA modules and DiT blocks while keeping the FFN modules unchanged. We also find that within MHA, attention redundancy varies across depth and between self- and cross-attention, so a uniform head ratio is inappropriate in few-step VDMs (Chavhan et al., 2026). We employ this variation to build more compact attention layers with MHA-CSS (column subset selection), which measures output-space head contributions, derives a module-specific budget from spectral energy, selects the retained heads by CSS-inspired backward elimination, and refits one scalar per retained head. For DiT blocks, existing importance scores based on local Cosine similarity mainly measure

the similarity of input and output feature, and neglect the timestep-dependent block influence or error propagation to the final generated latent when skipping one DiT block. We instead use endpoint Ablation Error: at one timestep we bypass a block, execute the remaining denoising trajectory, and compare the resulting final latent endpoint with the unablated reference. We form a low-risk candidate set independently at each of the four solver timesteps and physically remove only their intersection. Finally, recovery distillation with a simple loss function is applied to the compact model to restore performance.

Our main contributions are:

- We introduce MHA-CSS for module-specific attention-head pruning. For each self- and cross-attention module, MHA-CSS constructs Gram matrices from the projected outputs of individual attention heads, derives the number of retained heads from their cumulative spectral energy across solver timesteps, and selects the retained head identities through worst-timestep CSS-inspired backward elimination. Scalar refitting further adjusts the retained heads to reconstruct the original MHA output. Our analysis shows that cross-attention is generally more compressible than self-attention, supporting module-specific rather than uniform head budgets.
- We introduce endpoint Ablation Error for DiT blocks scoring and timestep-aware static DiT blocks pruning. At each solver timestep, we temporarily remove one block, execute the remaining solver steps normally, and measure the resulting relative error at the final generated latent endpoint. We then intersect the low-error candidate sets obtained at all timesteps to produce one static block mask shared across the complete sampling process.
- The pruning masks of attention heads and DiT blocks are derived solely from forward passes over calibration set and define a fixed compact architecture without runtime routing or cross-timestep feature caching. After recovery distillation with teacher-student distillation and Flow Matching losses, our model reduces the parameter count by 25.4% relative to the four-step Wan2.2-TI2V-5B teacher model, while nearly preserving its VBench-I2V total score for video generation.

## 2 Related Work

### 2.1 Few-Step Video Diffusion

Distribution Matching Distillation (DMD) (Yin et al., 2024b) trains a few-step generator by matching the teacher distribution rather than repeating the full multi-step sampling trajectory. DMD2 (Yin et al., 2024a) improves the stability of distribution matching and extends the formulation to multi-step generation. For video diffusion, Self-Forcing (Huang et al., 2026) uses rollout-based training to reduce the train–test mismatch in autoregressive generation. Few-step distillation reduces sampling time without automatically reducing the parameter count, weight-storage footprint, or the peak memory required by an individual denoiser evaluation. This distinction is important for deployment on memory-constrained edge devices and motivates structural compression as a direction orthogonal to step reduction.

### 2.2 Feature Caching and Dynamic Pruning

Feature caching accelerates video diffusion by reusing intermediate features along a sampling trajectory. MagCache (Ma et al., 2026b) models the evolution of residual magnitudes to identify reusable computations, whereas EasyCache (Zhou et al., 2025) uses a runtime-adaptive policy to reuse transformation vectors. These methods can reduce repeated computation without retraining the denoiser, but they retain the original parameter set and require runtime cache state. They therefore provide temporal or runtime acceleration rather than structural compression of the model itself.

Dynamic pruning changes the amount of computation executed for a particular input or timestep. DiP-GO (Zhu et al., 2024) uses backup connections and a plug-in pruner to select a subnet. Similarity-based Token Pruning (Zhang et al., 2025) removes redundant tokens during inference according to token similarity. PARE (Wang et al., 2026) jointly considers width and depth reduction and uses a timestep- and content-conditioned router to determine which computation is executed.

MobileWan (Ghafoorian et al., 2026) learns binary head gates and combines them with distillation for mobile video generation. Unlike feature caching, these dynamic pruning methods may require training gates, routers, or distillation objectives. More broadly, they reduce computation along an input-dependent trajectory, but do not by themselves guarantee a single parameter-smaller model with a fixed execution graph.

### 2.3 Structural Pruning for Diffusion Models

Structural pruning directly removes parameters to change the model architecture rather than only shortening its sampling trajectory. Diff-Pruning (Fang et al., 2023) aggregates Taylor-based importance signals across diffusion timesteps and retrains the selected structure. TinyFusion (Fang et al., 2025) learns binary masks for DiT depth selection followed by post-pruning optimization. PPCL (Ma et al., 2026a) detects contiguous layer redundancy for DiT-block pruning. NanoFLUX (Chavhan et al., 2026) combines progressive pruning, token downsampling, and distillation for a mobile text-to-image flow model. Video-oriented structural pruning methods address additional temporal constraints. VDMini (Wu et al., 2025a) separates shallow content processing from deeper motion processing and recovers a block pruned student with content- and motion-consistency objectives. NeoDragon (Karnewar et al., 2026) removes MMDiT blocks using a cosine-similarity-based importance score adapted from SANA 1.5 (Xie et al., 2025) and applies a two-stage distillation pipeline. MobileVD (Yahia et al., 2025) studies the effect of temporal block removal in one-step video diffusion, while TAMING (Wu et al., 2025b) examines the different sensitivities of MHA, DiT blocks, and FFN modules for mobile video generation. These methods demonstrate the need to account for content, motion, and module heterogeneity, but often rely on local similarity scores, pruning masks, and routers obtained through a training process.

## 3 Methodology

Our pruning framework first derives structural masks from calibration forward passes and then applies recovery distillation, as illustrated in Fig. 1. SEAE has three stages: module-specific head pruning for self- and cross-attention, timestep-consistent DiT block pruning, and recovery distillation of the compact model. Given a frozen few-step teacher $f_{\theta_T}$, a calibration set $\mathcal{D}_{\text{cal}}$, and solver timesteps $\mathcal{T} = \{t_1, \dots, t_4\}$, we seek a compact static student that reduces attention width and DiT depth while preserving the teacher's generative capabilities. We derive the structural masks from forward-only calibration, and then recover the pruned model with distillation:

$$\begin{aligned}(\mathcal{M}_{\text{MHA}}, \mathcal{M}_{\text{DiT}}) &= \text{Calibrate}(f_{\theta_T}, \mathcal{D}_{\text{cal}}, \mathcal{T}),\\ f_{\theta_S} &= \text{Prune}\big(f_{\theta_T}; \mathcal{M}_{\text{MHA}}, \mathcal{M}_{\text{DiT}}\big).\end{aligned} \tag{1}$$

Here, $\mathcal{M}_{\text{MHA}}$ contains the module-specific retained-head subsets and $\mathcal{M}_{\text{DiT}}$ contains the static block mask. Next, we introduce the Flow Matching preliminaries in Sec. 3.1, and then introduce MHA-CSS and DiT Blocks Pruning in Sec. 3.2 and Sec. 3.3, respectively. The recovery objective is defined in Sec. 3.4.

### 3.1 Preliminaries: Flow Matching

Flow Matching regresses a vector field to the velocity of a prescribed path (Lipman et al., 2023). For clean latent $x_0$, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, text-image condition $c$, and continuous level $s \in [0, 1]$, let $\alpha_s$ and $\sigma_s$ denote differentiable path schedules and let $v_\theta$ denote the predicted velocity field. We use

$$x_s = \alpha_s x_0 + \sigma_s \epsilon, \quad u_s = \dot{\alpha}_s x_0 + \dot{\sigma}_s \epsilon, \qquad \mathcal{L}_{\text{FM}} = \mathbb{E}\|v_\theta(x_s, s, c) - u_s\|_2^2. \tag{2}$$

For the rectified path, $\alpha_s = 1 - s$ and $\sigma_s = s$, so $x_s = (1 - s)x_0 + s\epsilon$ and $u_s = \epsilon - x_0$. Sampling follows $\mathrm{d}x_s/\mathrm{d}s = v_\theta(x_s, s, c)$ from the noise endpoint $x_1$ to the data endpoint as $s$ decreases from 1 to 0.

### 3.2 Calibration-only MHA-CSS for Module-Specific Attention Head Pruning

Attention redundancy varies across DiT layers and between self- and cross-attention, making a uniform head-retention ratio suboptimal. MHA-CSS therefore addresses two questions for each atten-

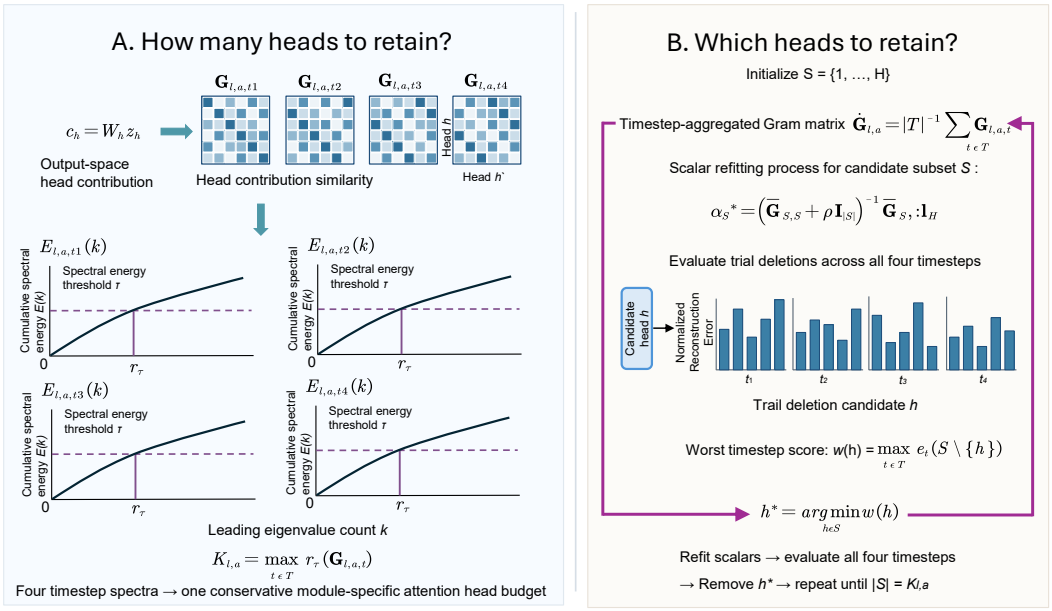


Figure 2: MHA-CSS determines a module-specific head budget from timestep-wise Gram spectra, then iteratively removes the head with the smallest worst-timestep reconstruction error until the budget is reached.

tion module: how many heads should be retained, and which heads should be retained. As illustrated in Fig. 2, we first derive a module-specific head budget from the cumulative spectral energy of projected attention-head outputs across solver timesteps. We then use CSS-inspired backward elimination with scalar refitting to select one retained-head subset that controls the worst reconstruction error across all timesteps.

For block $l \in \{1, \dots, L\}$, attention type $a \in \{\text{self}, \text{cross}\}$, and head $h \in \{1, \dots, H\}$, let $D$ be the residual width, $d$ the head width, and $\mathbf{z}_{l,a,h}(n) \in \mathbb{R}^d$ the output of head $h$ at token $n$. Let $\mathcal{T}$ denote the four solver timesteps and let $m$ index the $M$ calibration samples. Partitioning the output projection into head-sized blocks gives the output-space contribution

$$W^O_{l,a} = [W^O_{l,a,1}, \dots, W^O_{l,a,H}], \quad \mathbf{c}_{l,a,h}(n) = W^O_{l,a,h}\mathbf{z}_{l,a,h}(n). \tag{3}$$

Thus redundancy is measured after the learned output projection, rather than from attention activations alone. For sample $m$ and timestep $t \in \mathcal{T}$, define $C^{(m)}_{l,a,t}(n) = [\mathbf{c}^{(m)}_{l,a,t,1}(n), \dots, \mathbf{c}^{(m)}_{l,a,t,H}(n)] \in \mathbb{R}^{D\times H}$ and $s^{(m)}_{l,a,t,n} = \operatorname{tr}(C^{(m)\top}_{l,a,t}(n)C^{(m)}_{l,a,t}(n))$. We retain valid tokens with finite positive $s^{(m)}_{l,a,t,n}$, denote their set by $\mathcal{V}^{(m)}_{l,a,t}$, and write $N^{(m)}_{l,a,t} = |\mathcal{V}^{(m)}_{l,a,t}|$. Averaging trace-normalized token Grams and then giving each calibration sample equal weight gives

$$G^{(m)}_{l,a,t} = \frac{1}{N^{(m)}_{l,a,t}} \sum_{n\in\mathcal{V}^{(m)}_{l,a,t}} \frac{C^{(m)\top}_{l,a,t}(n)C^{(m)}_{l,a,t}(n)}{s^{(m)}_{l,a,t,n}}, \quad G_{l,a,t} = \frac{1}{M}\sum_{m=1}^{M} G^{(m)}_{l,a,t}. \tag{4}$$

Each layer and attention type has its own Gram statistics; no self-/cross-attention or layer statistics are shared. Let $\lambda_1 \geq \cdots \geq \lambda_H$ be the ordered eigenvalues of $G_{l,a,t}$. The effective spectral rank and retained-head budget are

$$r_\tau(G_{l,a,t}) = \min\left\{k : \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{H}\lambda_i} \geq \tau\right\}, \quad K_{l,a} = \max_{t\in\mathcal{T}} r_\tau(G_{l,a,t}). \tag{5}$$

Thus, $\tau$ is a cumulative spectral-energy threshold, while $K_{l,a}$ is the single module-specific head budget that covers all four timesteps.

The budget determines how many heads to retain, while CSS determines which ones. For a candidate retained set $S \subseteq \{1, \dots, H\}$ and $\bar{G} = |\mathcal{T}|^{-1}\sum_{t\in\mathcal{T}} G_t$, we refit one scalar per retained head, shared

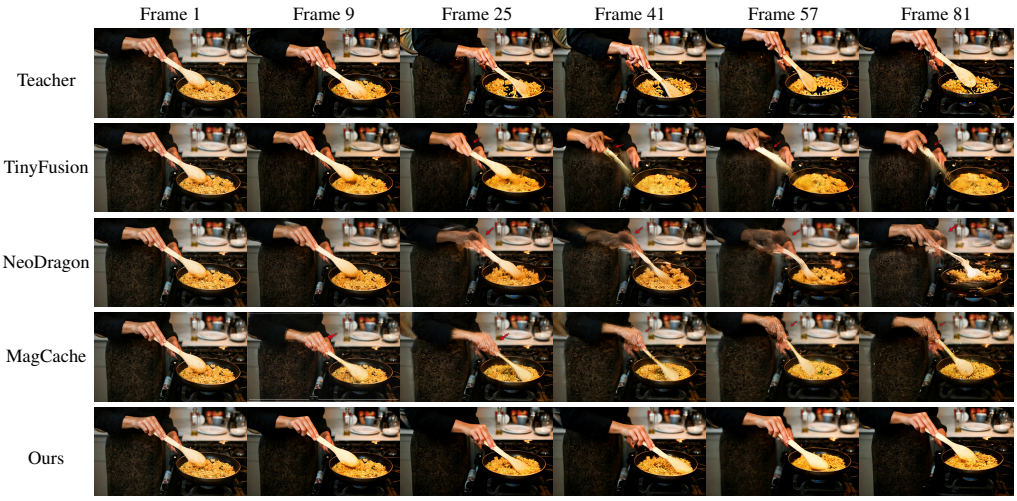


Figure 3: Qualitative comparison with video acceleration methods on TI2V sample. Prompt: *"a woman is stirring food in a pan on the stove"*.

across timesteps. The refit is the closed-form ridge minimizer of reconstructing the sum of all projected head contributions:

$$\boldsymbol{\alpha}_S^* = (\bar{G}_{S,S} + \rho I_{|S|})^{-1} \bar{G}_{S,:} \mathbf{1}_H, \qquad \boldsymbol{\alpha}_S \in \mathbb{R}^{|S|}, \quad \mathbf{1}_H \in \mathbb{R}^H. \tag{6}$$

Here $\bar{G}_{S,S}$ is the principal submatrix, $\bar{G}_{S,:}$ contains the rows indexed by $S$, $I_{|S|}$ is the $|S| \times |S|$ identity, $\mathbf{1}_H$ is the $H$-dimensional all-ones vector, and $\rho > 0$ is a ridge coefficient. For timestep $t$, let $G_t = G_{l,a,t}$ and define the normalized reconstruction error after refitting by

$$\begin{aligned} e_t(S) &= \frac{[E_t(S)]_+}{\max(d_t, \epsilon)}, \qquad d_t = \mathbf{1}_H^\top G_t \mathbf{1}_H, \\ E_t(S) &= d_t - 2\boldsymbol{\alpha}_S^{*\top} G_{t,S,:} \mathbf{1}_H + \boldsymbol{\alpha}_S^{*\top} G_{t,S,S} \boldsymbol{\alpha}_S^*, \qquad [x]_+ = \max(x, 0). \end{aligned} \tag{7}$$

During each tentative deletion, the scalar coefficients are refit before this error is evaluated. Backward CSS starts with all heads and repeatedly removes

$$h^* = \arg\min_{h \in S} \max_{t \in \mathcal{T}} e_t(S \setminus \{h\}) \quad \text{until} \quad |S| = K_{l,a}. \tag{8}$$

This pruning-mask construction uses calibration forward passes only. The selected Q/K/V rows, Q/K normalization channels, and output projection columns are physically sliced. CSS scales are folded into the retained output projection blocks, while the residual/FFN dimension is unchanged.

### 3.3 Timestep-Aware Ablation Error for Static DiT Block Pruning

After reducing the attention width, we prune redundancy along the DiT depth. Cosine similarity used by SANA1.5 (Xie et al., 2025) and NeoDragon (Karnewar et al., 2026), measures only the immediate input and output feature change caused by a block and do not capture how its removal propagates through the remaining few-step sampling process. We therefore evaluate each block by temporarily bypassing it at one solver timestep, completing all subsequent solver steps, and measuring the error at the final generated latent endpoint. The low-error candidate sets obtained at all timesteps are then intersected to produce one static block mask shared by the complete sampling process.

For calibration sample $m \in \{1, \ldots, M\}$, timestep $t \in \mathcal{T}$, and block $l$, we keep the condition, initial noise, scheduler, and all other operations fixed, replace only block $l$ by the identity at $t$, and then execute the remaining solver steps normally. Let $z_{m,\text{end}}^{\text{ref}}$ be the final latent of the unablated reference trajectory and $z_{m,\text{end}}^{(-l,t)}$ the endpoint after this single ablation. The generated latent region $\Omega_{\text{gen}}$ excludes the conditioning frame. We define the endpoint Ablation Error as

$$A_{t,l} = \frac{1}{M} \sum_{m=1}^{M} \frac{\|z_{m,\text{end}}^{(-l,t)} - z_{m,\text{end}}^{\text{ref}}\|_{2,\Omega_{\text{gen}}}^2}{\max(\|z_{m,\text{end}}^{\text{ref}}\|_{2,\Omega_{\text{gen}}}^2, \epsilon)}. \tag{9}$$

Table 1: Quantitative comparison of efficiency and VBench-I2V on the four-step Wan2.2-TI2V-5B Turbo (see full results in Table 4 in Appendix).

| Method | Params (B) ↓ | Efficiency | | | | VBench-I2V | | |
|---|---|---|---|---|---|---|---|---|
| | | FLOPs (P) ↓ | Speedup ↑ | DiT latency (s) ↓ | Peak VRAM (GiB) ↓ | Quality score ↑ | I2V score ↑ | Total score ↑ |
| Wan2.2-Turbo (Li et al., 2025) | 5 | 0.2913 | 1.00 | 1.20 | 21.21 | 0.7961 | 0.9477 | 0.8719 |
| TinyFusion (Fang et al., 2025) | 3.85 | <u>0.2244</u> | 1.20 | 0.92 | <u>19.06</u> | 0.7662 | 0.9558 | 0.8610 |
| NeoDragon (Karnewar et al., 2026) | 3.83 | <u>0.2244</u> | 1.20 | 0.92 | 19.06 | <u>0.7690</u> | 0.9548 | 0.8619 |
| MagCache (Ma et al., 2026b) | 5 | **0.1479** | **1.97** | **0.61** | 21.26 | 0.7671 | <u>0.9570</u> | <u>0.8621</u> |
| Ours | 3.73 | 0.2261 | <u>1.36</u> | <u>0.88</u> | **18.84** | **0.7805** | **0.9598** | **0.8702** |

The denominator normalizes each sample by the reference endpoint energy over $\Omega_{\text{gen}}$, which contains only the generated latent frames. Thus, a lower endpoint error means that removing the block has less downstream influence on the final generations. Let $K_t$ denote the candidate budget at timestep $t$:

$$\mathcal{C}_t = \text{BottomK}_l(A_{t,l}; K_t), \qquad \mathcal{S}_{\text{block}} = \bigcap_{t \in \mathcal{T}} \mathcal{C}_t. \tag{10}$$

### 3.4 Recovery Distillation Training

With the teacher $f_{\theta_T}$ frozen, recovery distillation trains the compact student $f_{\theta_S}$. The loss $\mathcal{L}_{\text{KD}}$ is a masked MSE between student and teacher predictions, while $\mathcal{L}_{\text{FM}}$ is a MSE between the student prediction and the Flow Matching target. The recovery objective is

$$\mathcal{L}_{\text{rec}} = \lambda \mathcal{L}_{\text{KD}} + (1 - \lambda)\mathcal{L}_{\text{FM}}, \qquad 0 \leq \lambda \leq 1. \tag{11}$$

Here, $\lambda$ controls the trade-off between output distillation and flow-matching supervision.

## 4 Experiments and Results

### 4.1 Experimental Setup

**Base Model.** We use the four-step Turbo model of Wan2.2-TI2V-5B as the teacher (Li et al., 2025; Wan et al., 2025) and the base model for our pruning experiments. The DiT backbone comprises 30 transformer blocks, each with separate self- and cross-attention modules. We prune the MHA modules and DiT blocks while keeping the FFN modules intact throughout the pruning process.

**Baseline.** For the text-image-to-video (TI2V) pruning task, we compare with structural pruning baselines TinyFusion (Fang et al., 2025) and NeoDragon (Karnewar et al., 2026), together with simple and random pruning baselines. We also include MagCache (Ma et al., 2026b), a cache-based video generation acceleration method, to cover a non-structural pruning alternative. All experiments are implemented on the four-step Wan2.2-TI2V-5B (Li et al., 2025) base model.

**Training Setup.** Our self-collected training set contains approximately 29,000 video clips, each with 81 frames at 16 FPS, a 5-second duration, and 480p resolution; every clip is paired with a text prompt. We randomly sample 500 representative text-video pairs for calibration. We evaluate the recovered model and other SOTA methods on the full VBench-I2V++ benchmark (Huang et al., 2025). We use VBench-I2V mini100, a random subset of VBench-I2V, to evaluate the pruned-init model in the ablation experiments. Appendix A.3 lists the full prompt information of Vbench-I2V mini100. All training, testing, and calibration experiments were conducted using the same GPU devices.

**Evaluation Metrics.** In addition to the ten VBench-I2V dimensions, we also report model parameters, training steps, denoising steps, FLOPs, speedup, DiT latency, and peak VRAM. These quantities measure the efficiency and quality trade-off of pruning. We define speedup as the teacher DiT latency divided by the student DiT latency. Other metrics difinitions can be found in Appendix A.1.

Table 2: Performance comparison of different spectral-energy threshold $\tau$.

| Threshold $\tau$ (%) | Self-/cross-attn heads | Params reduced (B) | VBench-I2V mini-100 ↑ |
|---|---|---|---|
| 95 | 640/510 | 0.423 | 0.868 |
| 90 | 563/449 | 0.673 | 0.850 |
| 85 | 506/383 | 0.867 | 0.825 |
| 80 | 451/328 | 1.040 | 0.737 |

Table 3: MHA-CSS component ablation under a fixed global head budget.

| Exp. | CSS | Scalar Refit | MHA Budget | VBench-I2V mini-100 ↑ |
|---|---|---|---|---|
| Ours | ✓ | ✓ | $\tau = 0.85$ | **0.825** |
| 2 | × | ✓ | $\tau = 0.85$ | 0.730 |
| 3 | × | ✓ | uniform | 0.721 |
| 4 | × | × | uniform | 0.671 |

### 4.2 Comparison with SOTA Methods

As illustrated in Fig. 3 and Fig. 7 (Appendix), our method preserves object structure and temporally consistent appearance more faithfully than the competing acceleration methods. Table 1 summarizes the corresponding efficiency and quality quantitative results. Our method's pruning rate reaches 25.4%, and obtains the highest total Vbench-I2V score among the accelerated methods listed in the table except the four-step teacher model Wan2.2-TI2V-5B Turbo. Our model achieves the VBench-I2V score closest to that of the teacher, while exhibiting lower DiT inference latency and peak VRAM usage. These results are measured after recovery training, and the pruning mask itself is obtained by forward-only calibration in our method, whereas TinyFusion and NeoDragon require training to obtain their pruning masks. MagCache is a training-free method and gives the best latency and speedup, but its quality scores are lower and its peak VRAM exceeds both ours and the teacher's peak VRAM. This memory profile may be less attractive for low-memory edge deployments. Overall, Table 1, Fig. 3 and Fig. 7 (Appendix) support a balanced efficiency-quality trade-off for our method rather than a universal dominance across every metric.

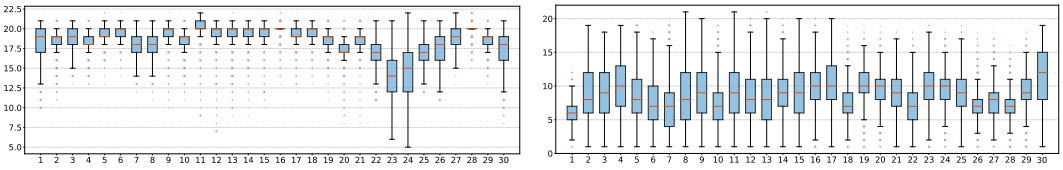

Figure 4: Layer-wise distributions of the effective spectral rank $r_{0.95}$ for self- and cross-attention. Outlier marker size scales with outlier frequency. Y-axis: 95% energy effective rank, X-axis: DiT layer index.

### 4.3 Ablation Studies

**Spectral Energy Threshold of MHA Pruning.** We display how spectral energy threshold $\tau$ influence the final quantitative and qualitative results in this section.

Fig. 4 characterizes how the pruning budget varies across modules, where $r_{0.95}$ is defined as the minimum integer rank required to capture 95% of the head-contribution spectral energy, and its maximum value is $H = 24$. This rank is neither the number of attention modules nor the final number of retained heads. As shown in Fig. 4, it is obvious that self-attention generally exhibits larger $r_{0.95}$ values than cross-attention, indicating a less compressible subspace. These per-module ranks determine the budgets $K_{l,a}$, after which MHA-CSS selects the actual heads. As shown in Fig. 5, the resulting budgets at $\tau = 0.85$ vary across layers and attention types, supporting module-specific rather than globally shared widths.

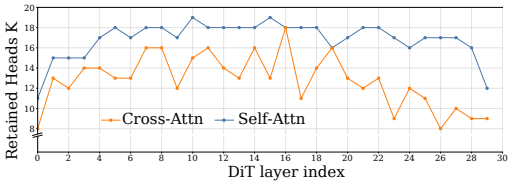


Figure 5: Module-specific head budgets $K_{l,a}$ obtained by MHA-CSS with $\tau = 0.85$.

Table 2 quantifies the trade-off controlled by the cumulative spectral-energy threshold $\tau$: lowering $\tau$ reduces the module-specific head budgets and increases parameter savings, but also degrades

generation quality. For $\tau \geq 0.85$, the pruned models largely preserve the VBench-I2V score no less than 0.8 without recovery training, whereas more aggressive pruning causes a sharp performance drop in Vbench-I2V score, indicating that excessive head removal harms representational capacity. The qualitative results in Fig. 8 (Appendix) support this trend: the $\tau = 0.95$ and $0.90$ variants remain visually close to the teacher without training, preserving the pouring stream and clear cup lettering. At $\tau = 0.80$, these structures degrade noticeably, in agreement with lower VBench-I2V scores in Table 2. After 1000 steps of recovery distillation training, $0.80$ variants recover most of the lost structure. These results from Table 2 and Fig. 8 (Appendix) motivate $\tau = 0.85$ as a practical operating point.

**Disentangling the Contributions of MHA-CSS.** We isolate the contributions of CSS-based head selection and scalar refitting with a component ablation under a fixed global budget, as shown in Table 3. Every variant retains 889 of the original 1,440 heads. With the same layer-wise budgets and scalar refitting procedure, replacing CSS by within-module random selection causes an absolute drop of 9.5% (11.52% relative) in Vbench-I2V, showing that the CSS selection rule matters for MHA pruning.

Scalar refitting provides a separate benefit: comparing the uniform and random variants shows an absolute gain of 0.05 in VBench-Total. The complete MHA-CSS pipeline consequently improves over the uniform and random variant without refitting by an absolute gain of 0.154.

**Ablation Error vs. Cosine Similarity.** Fig. 9 (Appendix) shows that the two selection criteria produce visibly different results. We ablate the two lowest-scoring blocks under each criterion and then use the same recovery training to generate the final videos. As illustrated in Fig. 9 (Appendix), in both examples, Cosine Similarity rows can accumulate structural perturbations and texture artifacts. The Ablation Error rows retain the main structure more consistently, which demonstrates that they selected less influential DiT blocks.

**High-Noise Steps Are More Sensitive to Block Ablation.** Fig. 6 clearly shows that block influence depends on the solver timestep. For each panel, a block is skipped only at the indicated timestep and evaluated by the generated latent endpoint relative MSE. Lower values denote safer pruning candidates. It is noticeable that high-noise steps are more sensitive to block ablation than low-noise steps, as the MSE sum at timestep 1000 is much higher than at timestep 625. Fig. 10 (Appendix) shows the qualitative differences between skipping the 10 least-important blocks measured by Ablation Error at the high-noise timestep 1000 and the low-noise timestep 625. It is clear that skipping blocks at high-noise steps can cause more image-quality degradation and texture artifacts than skipping blocks at low-noise steps, as supported by the quantitative results in Fig. 6. The rankings of block importance in different timesteps change, so a single score at one step is insufficient for static pruning. We instead form a low-error candidate set at every timestep and physically remove only the common intersection. This restricts pruning to blocks that are safe at high-noise timesteps while retaining more flexibility at less-sensitive low-noise timesteps.

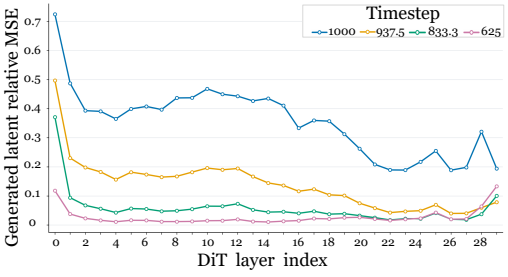


Figure 6: Timestep-wise DiT block Ablation Error.

## 5 Conclusion

We introduce SEAE, a structural pruning framework for VDMs whose pruning mask is obtained through forward-only calibration. SEAE first derives module-specific attention head budgets and applies CSS-based head selection with scalar refitting, and then ranks DiT blocks by their generated-latent endpoint Ablation Error at each solver timestep, physically removing blocks that are consistently associated with low loss across all timesteps. Without runtime routing or cross-timestep feature caching, SEAE reduces the parameter count by 25.4% while minimizing VBench-I2V performance degradation, demonstrating that substantial structural redundancy remains even in a few-step TI2V model.

## REFERENCES

Ruchika Chavhan, Malcolm Chadwick, Alberto Gil Couto Pimentel Ramos, Luca Morreale, Mehdi Noroozi, and Abhinav Mehrotra. Nanoflux: Distillation-driven compression of large text-to-image generation models for mobile devices. In *Forty-third International Conference on Machine Learning*, 2026.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Gongfan Fang, Kunjun Li, Xinyin Ma, and Xinchao Wang. Tinyfusion: Diffusion transformers learned shallow. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18144–18154. IEEE, 2025.

Mohsen Ghafoorian, Denis Korzhenkov, Adil Karjauv, Ioannis Lelekas, Noor Fathima, Spyridon Stasis, Hanno Ackermann, Boris van Breugel, Markus Nagel, Fatih Porikli, et al. Mobilewan: Closing the quality gap for mobile video diffusion. *arXiv preprint arXiv:2607.06173*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Animesh Karnewar, Denis Korzhenkov, Ioannis Lelekas, Noor Fathima, Adil Karjauv, Mohsen Ghafoorian, and Amirhossein Habibian. Neodragon: Mobile video generation using diffusion transformer. In *International Conference on Learning Representations*, volume 2026, pp. 39506–39537, 2026.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12112–12123. IEEE, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Jian Ma, Qirong Peng, Xujie Zhu, Peixing Xie, Chen Chen, and Haonan Lu. Pluggable pruning with contiguous layer distillation for diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18640–18650, 2026a.

Zehong Ma, Longhui Wei, Feng Wang, Shiliang Zhang, and Qi Tian. Magcache: Fast video generation with magnitude-aware cache. *Advances in Neural Information Processing Systems*, 38: 34348–34380, 2026b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yutong Wang, Yunke Wang, Tianfan Xue, Yu Qiao, Yaohui Wang, Xinyuan Chen, and Chang Xu. Pare: Pruning and adaptive routing for efficient video generation. *arXiv preprint arXiv:2605.27336*, 2026.

Yiming Wu, Zhenghao Chen, Huan Wang, and Dong Xu. Individual content and motion dynamics preserved pruning for video diffusion models. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pp. 9714–9723. ACM, 2025a.

Yushu Wu, Yanyu Li, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ke Ma, Arpit Sahni, Ju Hu, Aliaksandr Siarohin, Dhritiman Sagar, et al. Taming diffusion transformer for efficient mobile video generation in seconds. *arXiv preprint arXiv:2507.13343*, 2025b.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Haitam Ben Yahia, Denis Korzhenkov, Ioannis Lelekas, Amir Ghodrati, and Amirhossein Habibian. Mobile video diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19450–19460. IEEE, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Evelyn Zhang, Jiayi Tang, Xuefei Ning, and Linfeng Zhang. Training-free and hardware-friendly acceleration for diffusion models via similarity-based token pruning. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 9878–9886. AAAI Press, 2025.

Yang Zhang, Er Jin, Wenzhong Liang, Yanfei Dong, Ashkan Khakzar, Philip Torr, Johannes Stegmaier, and Kenji Kawaguchi. Learnable sparsity for vision generative models. In *International Conference on Learning Representations*, 2026.

Xin Zhou, Dingkang Liang, Kaijin Chen, Tianrui Feng, Xiwu Chen, Hongkai Lin, Yikang Ding, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. Less is enough: Training-free video diffusion acceleration via runtime-adaptive caching. *arXiv preprint arXiv:2507.02860*, 2025.

Haowei Zhu, Dehua Tang, Ji Liu, Mingjie Lu, Jintu Zheng, Jinzhang Peng, Dong Li, Yu Wang, Fan Jiang, Lu Tian, et al. Dip-go: A diffusion pruner via few-step gradient optimization. *Advances in Neural Information Processing Systems*, 37:92581–92604, 2024.

# A APPENDIX

The appendix will contain additional implementation details.




Figure 7: Qualitative comparison with video acceleration methods on TI2V sample. Prompt: *"a yellow boat is cruising in front of a bridge"*.

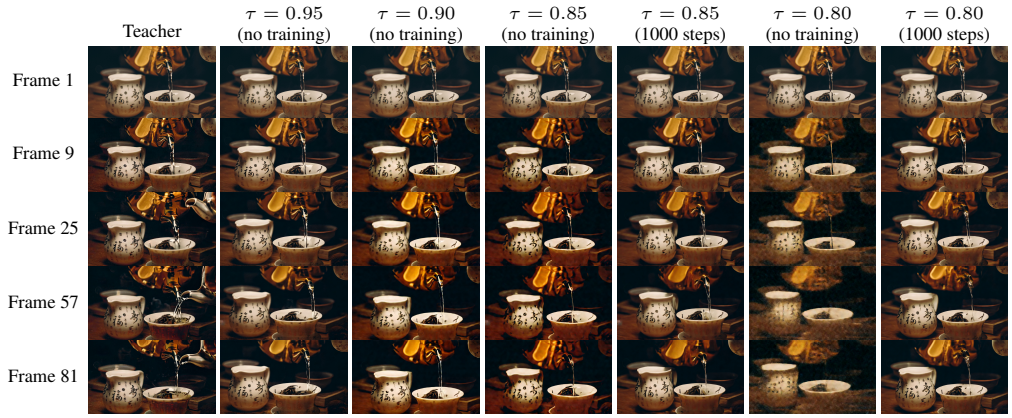


Figure 8: Qualitative comparison of the teacher and MHA-pruned variants at the indicated video frames. Prompt: *"a person is pouring water into a teacup"*.

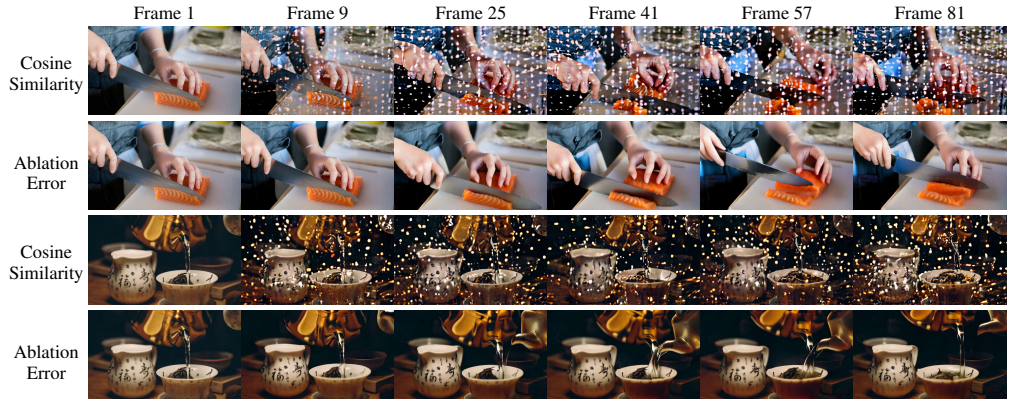


Figure 9: Qualitative comparison of two block selection criteria. Prompt A: *"a person slicing salmon on a cutting board"*, Prompt B: *"a person is pouring water into a teacup"*.

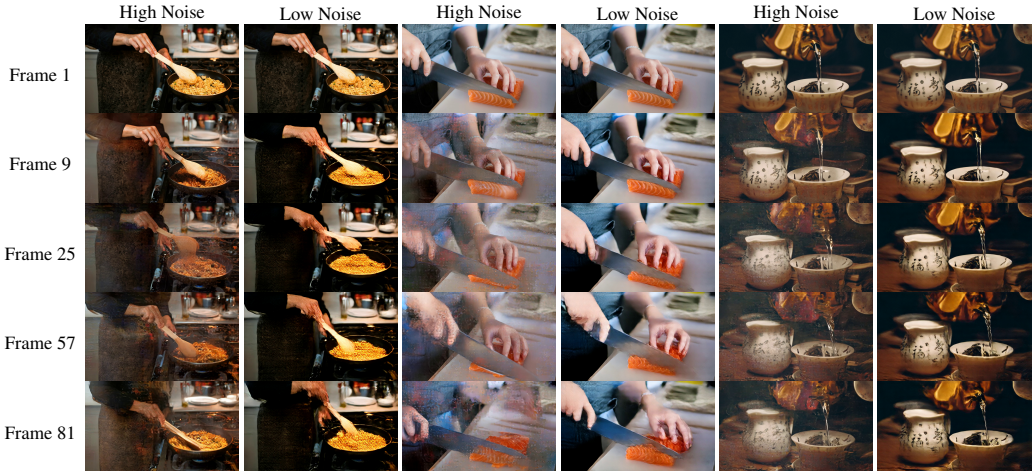


Figure 10: Qualitative comparison of skipping the 10 least-important blocks under high- and low-noise settings. Prompt A: *"a woman is stirring food in a pan on the stove"*, Prompt B: *"a person slicing salmon on a cutting board"*, Prompt C: *"a person is pouring water into a teacup"*.

### A.1 Evaluation Metrics

For completeness, this subsection describes the metrics reported in the main-text tables. The VBench-I2V score is reported along ten dimensions, each ranging over the benchmark's normalized score range (higher is better): *I2V subject* measures how well the generated subject preserves the subject in the input image; *I2V background* measures preservation of the input background; *camera motion* evaluates the quality and appropriateness of the generated camera motion; *subject consistency* measures whether the subject remains consistent across frames; *background consistency* measures temporal consistency of the background; *motion smoothness* measures the absence of undesirable inter-frame motion jitter; *dynamic degree* measures the amount of meaningful motion in the generated video; *aesthetic quality* evaluates the overall visual appeal; *imaging quality* evaluates perceptual properties such as sharpness and visual fidelity; and *temporal flickering* measures unwanted brightness, texture, or appearance changes over time. The tables additionally report the benchmark-provided aggregate *Quality score*, *I2V score*, and *Total score*, all of which are higher-is-better measures.

The efficiency metrics characterize the cost of the evaluated model and sampling configuration. *Params (B)* is the number of model parameters in billions; *recovery training steps* is the number of optimization steps used for recovery distillation after pruning; and *denoising steps* is the number of denoiser evaluations used to generate each video. *FLOPs (P)* reports the corresponding inference computation in petaflops, while *DiT latency (s)* is the measured latency of the DiT denoiser in seconds. *Peak VRAM (GiB)* is the maximum GPU memory occupied during inference. We define *speedup* as the teacher DiT latency divided by the student DiT latency. Thus, lower values are preferred for parameters, recovery training steps, denoising steps, FLOPs, DiT latency, and peak VRAM, whereas a higher speedup is preferred.

### A.2 Full Efficiency and VBench-I2V Results

Table 4: Quantitative comparison of efficiency and full VBench-I2V on the four-step Wan2.2-TI2V-5B Turbo.

| Method | Params (B) ↓ | Recovery Training steps ↓ | Denoising steps ↓ | Efficiency | | | |
|---|---|---|---|---|---|---|---|
| | | | | FLOPs (P) ↓ | Speedup ↑ | DiT latency (s) ↓ | Peak VRAM (GiB) ↓ |
| Wan2.2-Turbo (Li et al., 2025) | 5 | – | 4 | 0.2913 | 1.00 | 1.20 | 21.21 |
| TinyFusion (Fang et al., 2025) | 3.85 | 1,000 | 4 | 0.2244 | 1.20 | 0.92 | 19.06 |
| NeoDragon (Karnewar et al., 2026) | 3.83 | 1,000 | 4 | 0.2244 | 1.20 | 0.92 | 19.06 |
| MagCache (Ma et al., 2026b) | 5 | 0 | 2 | **0.1479** | **1.97** | **0.61** | 21.26 |
| Ours | 3.73 | 1,000 | 4 | 0.2261 | 1.36 | 0.88 | **18.84** |

| Method | I2V subject ↑ | I2V background ↑ | Camera motion ↑ | Subject consistency ↑ | Background consistency ↑ | Motion smooth. ↑ |
|---|---|---|---|---|---|---|
| Wan2.2-Turbo (Li et al., 2025) | 0.9812 | 0.9825 | 0.3596 | 0.9423 | 0.9450 | 0.9759 |
| TinyFusion (Fang et al., 2025) | 0.9882 | 0.9908 | 0.2810 | 0.9574 | 0.9680 | **0.9890** |
| NeoDragon (Karnewar et al., 2026) | 0.9885 | 0.9919 | 0.2944 | 0.9603 | 0.9653 | 0.9818 |
| MagCache (Ma et al., 2026b) | 0.9881 | 0.9910 | 0.3581 | 0.9329 | 0.9480 | **0.9890** |
| Ours | **0.9898** | **0.9928** | **0.3730** | **0.9630** | **0.9702** | 0.9821 |

| Method | Dynamic degree ↑ | Aesthetic quality ↑ | Imaging quality ↑ | Temporal flickering ↑ | Quality score ↑ | I2V score ↑ | Total score ↑ |
|---|---|---|---|---|---|---|---|
| Wan2.2-Turbo (Li et al., 2025) | 0.5276 | 0.6248 | 0.7059 | 0.9579 | 0.7961 | 0.9477 | 0.8719 |
| TinyFusion (Fang et al., 2025) | **0.3902** | 0.6233 | 0.6937 | **0.9846** | 0.7662 | 0.9558 | 0.8610 |
| NeoDragon (Karnewar et al., 2026) | 0.0699 | 0.6354 | **0.7063** | 0.9721 | 0.7690 | 0.9548 | 0.8619 |
| MagCache (Ma et al., 2026b) | 0.2000 | 0.6055 | 0.6920 | 0.9754 | 0.7671 | 0.9570 | 0.8621 |
| Ours | 0.1537 | **0.6503** | 0.7021 | 0.9694 | **0.7805** | **0.9598** | **0.8702** |

### A.3 VBench Prompts

The following 100 text prompts are used for the VBench evaluation results reported in our ablation experiments of Table 2 and Table 3. The corresponding input images are omitted for brevity.

1. *a bar with chairs and a television on the wall, camera pans right*

2. *a beach with a lot of buildings on the side of a cliff*
3. *A black and white abstract video featuring mesmerizing bubbles*
4. *a blue train traveling through a lush green area*
5. *a book on fire with flames coming out of it*
6. *a bridge that is in the middle of a river*
7. *a bridge that is over a body of water, camera tilts down*
8. *a bunch of food is cooking on a grill over an open fire*
9. *a bunch of houses that are on a hillside*
10. *a castle on top of a hill covered in snow*
11. *a city street at night during a snow storm, camera tilts up*
12. *a close up of a blue and orange liquid, camera zooms out*
13. *a close up of a piece of sushi on chopsticks*
14. *a close-up of a hippopotamus eating grass in a field*
15. *a clown fish hiding in a purple anemone*
16. *a dog carrying a soccer ball in its mouth*
17. *a foggy field that has trees in the grass, camera tilts up*
18. *a foggy road with trees in the distance*
19. *a group of camels resting in the desert*
20. *a group of hot air balloons flying over a field*
21. *a group of people gathered around a red checkered blanket*
22. *a house with snow on the ground, camera pans right*
23. *a kitchen with a sink and a lot of glasses on the counter*
24. *a lake surrounded by mountains and flowers*
25. *a large bonfire is burning in the night sky*
26. *a large pot of soup filled with vegetables and meat*
27. *a large rhino grazing in the grass near a bush*
28. *a large waterfall in the middle of a building, camera zooms out*
29. *a large waterfall in the middle of a lush green hillside*
30. *a large wave crashes into a lighthouse on a stormy day, camera tilts down*
31. *a large wave in the ocean with a lot of spray coming from it, camera pans left*
32. *a light house on the edge of the water*
33. *a little boy jumping in the air over a puddle of water*
34. *a little girl wearing a purple helmet riding a blue bike*
35. *a living room filled with furniture next to a stone wall*
36. *a living room filled with lots of books on a wall, camera tilts down*
37. *a living room with hardwood floors and a white couch, camera zooms in*
38. *A man and woman dressed as sugar skulls in a field of flowers, sharing a loving gaze with each other*
39. *a man holding a baby in his arms*
40. *a man in sunglasses laying on a wooden bench*
41. *a man sitting on steps playing an acoustic guitar*
42. *a man standing on top of a mountain with a backpack*
43. *a monastery sits on top of a cliff in bhutan, camera static*
44. *a mountain covered in snow with evergreen trees*

45. *a mountain range with snow on the ground, camera static*
46. *a mountain with snow on it*
47. *a painting of a cloudy sky next to an easel*
48. *a person holding a sparkler in front of a sunset*
49. *A person is making dumplings*
50. *a person is putting sauce on a burger*
51. *a pile of wood sitting next to a log house, camera tilts down*
52. *a pink lotus flower in the middle of a pond*
53. *a plant with water droplets on it*
54. *a purple and yellow abstract painting with a black background, camera zooms out*
55. *a reflection of a city with buildings in the water*
56. *a room filled with books and teddy bears, camera zooms out*
57. *a room filled with paintings and statues*
58. *a sandy beach with palm trees on the shore, camera pans left*
59. *a small monkey holding a piece of food in it's mouth*
60. *a snow covered street*
61. *a space shuttle taking off into the sky*
62. *A space station orbited above the Earth*
63. *a squirrel sitting on the ground eating a piece of bread*
64. *a table and chairs in a room with a plant in the corner*
65. *a very large waterfall in the middle of the day*
66. *a view of a snowy mountain side with many buildings, camera pans left*
67. *a visually captivating abstract video, rich in color, set against a dramatic black background*
68. *a waterfall is shown in the middle of a lush green hillside, camera static*
69. *a woman is praying in front of a buddhist temple*
70. *a woman sitting on the ground in front of a guitar*
71. *a woman walking down a sidewalk with a bag*
72. *a woman with afro hair is smiling while wearing earphones*
73. *a woman with curly hair is drinking a beer*
74. *a wooden table in front of a brick wall with bottles on the wall*
75. *A yellow boat is cruising in front of a bridge*
76. *a young man is covered in colored powder*
77. *a zebra walking across a dirt road near a field*
78. *an aerial view of a body of water and a beach*
79. *an aerial view of a curvy road in the middle of a forest, camera tilts up*
80. *an aerial view of a group of islands in the middle of a lake*
81. *an aerial view of a rocky beach in indonesia, camera pans right*
82. *an alley way in an old european city*
83. *an art gallery with paintings on the walls, camera zooms in*
84. *an asian city street at night with people and bicycles, camera tilts up*
85. *an older woman and a young girl are knitting together*
86. *an orange mushroom sitting on top of a tree stump in the woods*
87. *cars parked on the side of the road during a snowstorm, camera static*
88. *Chopsticks are picking up noodles from the bowl*

89. *Colorful buildings on the seaside cliffs*
90. *reeds blowing in the wind against a cloudy sky*
91. *several hot air balloons flying over a city*
92. *sun rays shining through clouds over a lake, camera zooms in*
93. *the dome of the rock in jerusalem, camera zooms in*
94. *the golden gate bridge in san franscisco is lit up by the setting sun, camera pans left*
95. *the great wall of china in autumn*
96. *the pyramids of giza, egypt*
97. *the sun is setting over a lake in the mountains, camera pans right*
98. *the town of hallstatt is surrounded by mountains and water*
99. *tower bridge in london*
100. *two women eating pizza at a restaurant*