Content-Aware Dynamic Patchification for Efficient Video Diffusion

Supplementary Material

6. Evaluation on Higher Resolution

In this section, we evaluate the performance of our proposed DynaPatch framework when generating higher resolution videos at 720p. As shown in Table 3, DynaPatch maintains VBench Total Scores close to the 720p baseline, which always uses $(1, 2, 2)$ patch size, while providing notable speedups. For example, at a 30% token reduction rate, DynaPatch achieves a Total Score of 82.87 with a 1.5× speedup, compared to the baseline score of 83.17. Compared with other patchification and token pruning approaches, DynaPatch consistently delivers higher VBench scores under the same token reduction rates. At 30% token reduction, FlexiDiT, D^{2}iT, and SPViT obtain Total Scores of 79.42, 80.16, and 78.73, respectively, with similar or lower speedups than DynaPatch. These results indicate that DynaPatch generalizes well to higher resolutions and can effectively perform content-aware dynamic patchification even when the number of spatial tokens grows significantly, maintaining strong visual quality while preserving computational gains.

Table 3. Evaluation on higher resolution (720p) videos.

Token Reduction	Method	Total Score ↑	Speedup ↑
	Baseline	83.17	1.0x
20%	FlexiDiT [1]	81.08	1.3x
	D^{2}iT [12]	81.26	1.2x
	SPViT [14]	80.13	1.3x
	DynaPatch (Ours)	83.10	1.3x
30%	FlexiDiT [1]	79.42	1.5x
	D^{2}iT [12]	80.16	1.5x
	SPViT [14]	78.73	1.4x
	DynaPatch (Ours)	82.87	1.5x
40%	FlexiDiT [1]	78.96	1.8x
	D^{2}iT [12]	78.42	1.7x
	SPViT [14]	77.03	1.8x
	DynaPatch (Ours)	81.31	1.8x

7. Evaluation on FVD Metric

In addition to VBench, we further compare our proposed DynaPatch with different patchification and token pruning methods using the Fréchet Video Distance (FVD) metric [28]. We conduct this experiment on the Adobe internal video dataset by sampling 128 prompts, and for each prompt, we first generate a reference video using the baseline model that always adopts the finest patch size $(1, 2, 2)$ (i.e., without dynamic patchification). Then, for each patchification and pruning method and token reduction rate, we generate videos under the same prompts and sampling configuration, and compute FVD between the method's outputs and the baseline outputs. As shown in Table 4, compared to FlexiDiT, D^{2}iT, and SPViT, DynaPatch achieves the lowest FVD at all token reduction rates while providing comparable speedups. For example, at a 30% token reduction rate, DynaPatch attains an FVD of 445.6 with a 1.5× speedup, whereas FlexiDiT, D^{2}iT, and SPViT obtain 491.5, 481.2, and 513.3, respectively. These results further demonstrate the effectiveness of the proposed DynaPatch framework in preserving video quality while providing speedups.

Table 4. Comparison of different methods using the FVD metric.

Token Reduction	Method	Speedup ↑	FVD ↓
20%	FlexiDiT [1]	1.3x	478.9
	D^{2}iT [12]	1.2x	472.7
	SPViT [14]	1.3x	497.4
	DynaPatch (Ours)	1.3x	428.3
30%	FlexiDiT [1]	1.5x	491.5
	D^{2}iT [12]	1.4x	481.2
	SPViT [14]	1.5x	513.3
	DynaPatch (Ours)	1.5x	445.6
40%	FlexiDiT [1]	1.8x	532.6
	D^{2}iT [12]	1.7x	517.9
	SPViT [14]	1.7x	559.1
	DynaPatch (Ours)	1.8x	478.8

8. Evaluation on Wan Model

In this section, we extend our evaluation to the public Wan2.1 model and the Mixkit dataset. As shown in Table 5, DynaPatch achieves 1.5× speedup with negligible quality drop, indicating the generalization ability of DynaPatch. Besides, initializing the Wan2.1 router with weights trained on our internal model accelerates convergence by ∼4×. This shows that the learned routing mechanism captures universal visual complexity cues rather than overfitting to a specific DiT.

Table 5. Evaluation on Wan2.1-1.3B finetuned on Mixkit dataset under two settings: *Transferred* (router initialized with weights learned from internal model) and *train from scratch*.

Token Reduct.	Method	VBench Total Score ↑ @ 5k steps	VBench Total Score ↑ @ 20k steps	Speedup↑
—	Baseline	83.31		1.0×
30%	DynaPatch (Transferred)	83.20	83.22	1.5×
	DynaPatch (Scratch)	79.10	83.17	
40%	DynaPatch (Transferred)	82.61	82.69	1.7×
	DynaPatch (Scratch)	77.55	82.48	

9. Training stability and router selection behavior analysis

Figure 7 shows the training loss curve and the router selection behavior. The loss curve shows a rapid initial drop and subsequent stability, which validates a stable training. The selection transits from near-uniform to a budget-respecting distribution in ∼5k steps, then stays relatively stable but fluctuates adaptively to optimize content-aware decisions. At 30% reduction, the average ratios are 58% (1,2,2), 38% (1,4,4), 4% (2,2,2). The rare (2,2,2) selection reflects learned preference for motion fidelity, while the trade-off between (1,2,2) and (1,4,4) shows dynamic adaptation to spatial complexity.

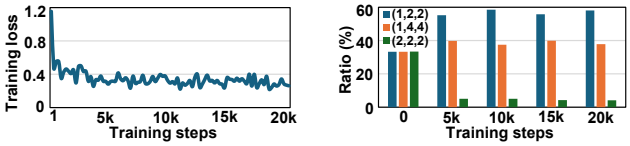


Figure 7. Loss curve and patch select ratios (30% token reduction).

10. Ablation Study on Patch Size Candidates

In our design, the router chooses among three patch sizes, $(1, 2, 2)$, $(2, 2, 2)$, and $(1, 4, 4)$ for each region. To test whether a larger candidate set can further improve performance, we extend the pool by adding three more patch sizes, $(1, 8, 8)$, $(2, 4, 4)$, and $(2, 8, 8)$, and compare the original three-size configuration with this six-size configuration. As shown in Table 6, introducing additional patch sizes does not yield higher gains in either VBench score or speedup, and can even slightly hurt performance. For example, at a 40% token reduction rate, the Total Score decreases from 82.19 with three patch sizes to 81.75 with six patch sizes. These results suggest that our default configuration of three patch sizes is already sufficient to effectively exploit both spatial and temporal redundancy. Adding more candidates could increase the complexity of routing decisions without providing extra benefits. Moreover, in practice, we observe that the newly added coarse patch sizes are rarely selected: their combined selection rate is below 2%, as they are too aggressive and tend to noticeably degrade the quality of the generated videos.

11. Detailed Results on all Dimensions of VBench

In this section, we report the full VBench scores for all evaluation dimensions, corresponding to the results in Table 1 in Section 4.2 in the main paper. Table 7 lists the scores on quality-related dimensions, while Table 8 shows the scores on semantic-related dimensions.

Table 6. Ablation study on patch size candidates. Ablation study on patch size candidates Ablation study on patch size candidates.

Token Reduction	Method	Total Score ↑	Speedup ↑
	Baseline	83.61	1.0x
20%	Base (three patch sizes)	83.56	1.3x
	Six Patch Sizes	83.58	1.3x
30%	Base (three patch sizes)	83.42	1.5x
	Six Patch Sizes	83.30	1.5x
40%	Base (three patch sizes)	82.19	1.8x
	Six Patch Sizes	81.75	1.8x

12. User study

Besides the quantitative results, we also conducted a blind user study (with 10 humans, 50 videos, 1-10 scale on two metrics: Semantic Alignment and Visual Quality). In our user study, DynaPatch achieves comparable ratings to baseline without token reduction: alignment 9.2 (vs. 9.3 baseline), quality 8.5 (vs. 8.6 baseline). These results indicate a negligible impact on human-perceived quality when using our proposed DynaPatch.

13. Generated Video Visualization

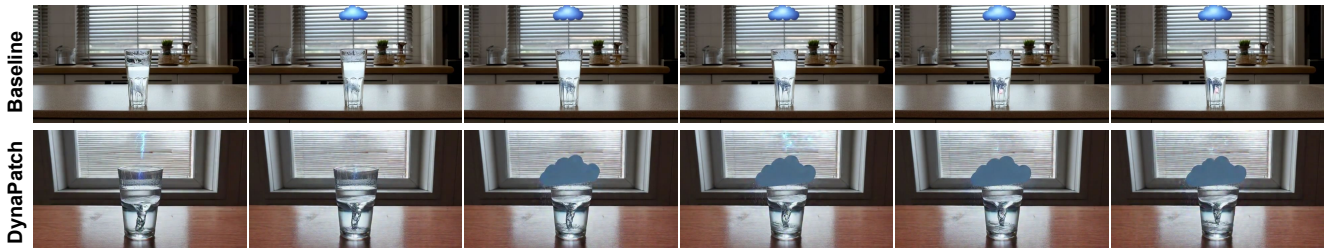
Because of space constraints in the main paper, we provide additional visualizations of generated videos in this section (Figure 8 and Figure 9), comparing the baseline that always uses the $(1, 2, 2)$ patch size and our DynaPatch with a 30% token reduction rate. As shown in the figures, DynaPatch effectively preserves temporal coherence, spatial structure, and semantic alignment with the text prompts, while maintaining visual quality close to the baseline despite reducing a significant fraction of tokens.

Table 7. Full VBench quality scores across all dimensions for FlexiDiT, D^{2}iT, SPViT, and our DynaPatch under different token reduction rates.

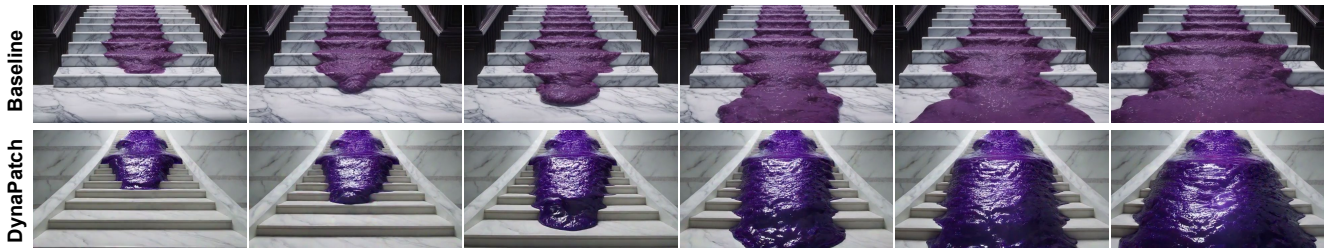
Token Reduction	Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Aesthetic Quality	Imaging Quality	Dynamic Degree
	Baseline	93.72	96.77	96.18	98.75	62.46	65.38	98.61
20%	FlexiDiT [1]	92.29	95.47	95.54	98.30	60.58	63.45	98.06
	D^{2}iT [12]	92.37	95.65	95.63	98.29	60.84	63.92	98.06
	SPViT [14]	92.15	95.36	95.41	98.19	60.09	63.37	97.78
	DynaPatch (Ours)	93.65	96.70	96.15	98.72	62.37	65.29	98.61
30%	FlexiDiT [1]	91.25	94.73	95.12	97.96	58.63	62.36	97.50
	D^{2}iT [12]	92.11	95.16	95.43	98.21	59.87	63.12	97.78
	SPViT [14]	90.80	94.06	94.92	97.73	57.52	61.11	97.22
	DynaPatch (Ours)	93.56	96.61	96.11	98.68	62.21	65.28	98.33
40%	FlexiDiT [1]	91.33	94.64	94.98	97.86	59.00	62.26	97.50
	D^{2}iT [12]	89.74	92.98	94.58	97.39	56.76	59.71	96.94
	SPViT [14]	90.12	93.12	94.39	97.52	56.87	60.38	96.67
	DynaPatch (Ours)	92.89	96.05	95.86	98.50	61.26	64.31	98.06

Table 8. Full VBench semantic scores across all dimensions for FlexiDiT, D^{2}iT, SPViT, and our DynaPatch under different token reduction rates.

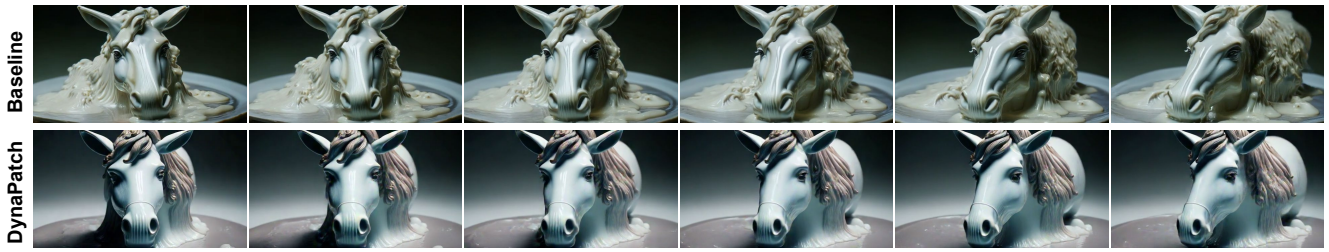
Token Reduction	Method	Object Class	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
	Baseline	95.09	77.36	98.00	90.25	67.97	46.80	22.42	25.21	26.92
20%	FlexiDiT [1]	92.74	75.53	94.80	87.06	65.40	44.93	21.69	24.43	26.12
	D^{2}iT [12]	90.60	73.55	94.40	86.79	66.28	45.30	21.59	24.27	25.74
	SPViT [14]	89.25	72.70	93.40	85.37	64.97	44.48	21.36	23.54	25.49
	DynaPatch (Ours)	94.57	77.76	97.80	90.88	68.05	46.88	22.46	25.28	26.73
30%	FlexiDiT [1]	87.97	72.60	91.40	83.12	63.79	44.16	21.06	23.46	24.75
	D^{2}iT [12]	89.43	72.79	91.80	84.18	64.87	44.12	21.21	23.65	25.58
	SPViT [14]	85.94	70.58	88.20	81.68	61.69	41.96	20.26	22.94	24.60
	DynaPatch (Ours)	95.31	76.49	98.00	90.44	67.80	46.94	22.26	24.99	26.79
40%	FlexiDiT [1]	84.99	70.13	87.80	80.79	61.50	42.17	20.10	22.47	24.34
	D^{2}iT [12]	87.00	70.62	89.40	82.23	62.66	42.27	20.31	22.68	24.12
	SPViT [14]	85.68	69.07	89.00	80.94	61.27	42.23	19.95	22.31	24.11
	DynaPatch (Ours)	91.61	74.95	93.80	86.21	64.58	44.55	21.42	24.07	25.90



Prompt: A glass of water on a Scandinavian kitchen counter. Natural light from blinds behind the main scene. There's a small storm cloud, about 40cm wide, over the glass of water, with thunder and rain pouring over the glass. The water inside the glass becomes a maelstrom eventually.

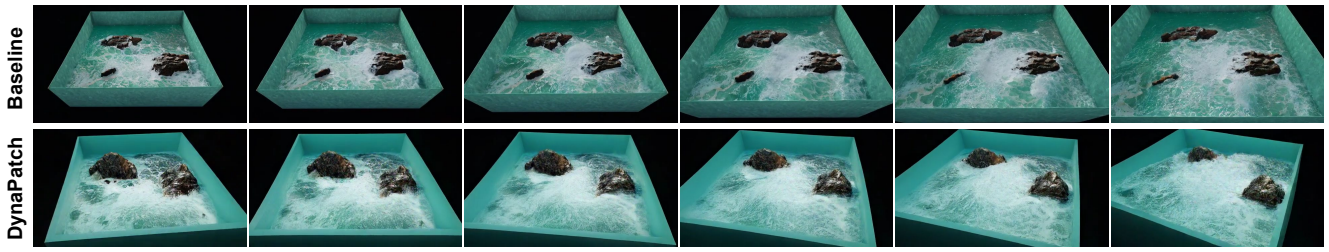


Prompt: A low angle hyper realistic shot of a thick purple goo flowing quickly down a white marble staircase. Cinematic, highly detailed, film grade.



Prompt: A highly detailed portrait of a porcelain donkey being covered by thick slime. Cinematic, highly detailed, film grade.

Prompt: A steam-powered alchemist in a Victorian-era lab with intricate machinery and bubbling potions. Mechanical limbs and tools whir and clink, blending old-world craftsmanship and mechanical innovation in a world where science and magic intertwine.

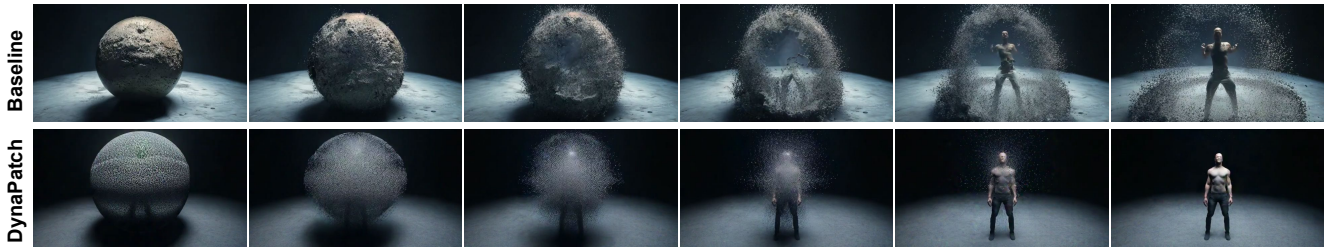


Prompt: A dynamic motion shot of a hyper-realistic ocean simulation confined to a 3D open box floating in darkness. Waves surge and recede, crashing against rocky outcrops with lifelike physics. Foam forms intricate patterns as water swirls and eddies around the stones. The camera slowly pans, capturing the play of light on the water's surface and the depth of the turquoise liquid. The open box emphasize the contrast between the vivid simulation and the surrounding black void.

Figure 8. Visualizations of the generated video by the baseline model and our DynaPatch framework.

Prompt: Generate a scene where a man touches a plant, causing it to transform into living vines that wrap around his body while he tries to remove them.

Prompt: A rabbit made of liquid gold with a dark background, liquidify, cinematic.



Prompt: Hyperspeed hand held camera. An irregular sphere shape ball dramatically undulates, warps and explodes as it transforms into a completely different man. Surreal.

Prompt: mystical video of a cat dressed as astrologer performing colorful magic spells in front of an eager crowd of village mice, the mice watch in awe as the cat conjures sparkling constellations, floating orbs, and magical symbols in the air.

Prompt: A futuristic city at twilight, with a vast hologram field in a central plaza. Holographic butterflies flutter among shifting geometric shapes, leaving luminescent trails. Focus on one vibrant butterfly with electric blue, purple, and silver wings, its delicate flight set against a backdrop of neon-lit skyscrapers and a stunning sunset. Ambient sounds of distant flying cars and melodic chimes enhance the serene yet energetic atmosphere.

Figure 9. Visualizations of the generated video by the baseline model and our DynaPatch framework (cont.).