

ReHyAt: Recurrent Hybrid Attention for Video Diffusion Transformers

Supplementary Material

1. Appendix

1.1. Training Details and Hyperparameters

Unless stated otherwise in the ablation studies, we parameterize ϕ using a two-layer MLP with a polynomial degree of 2. For each hybrid block, we apply separate transformations for keys and queries, denoted as ϕ_k and ϕ_q .

Pretraining (Distillation Stage). During pretraining, each block is trained independently while all parameters remain frozen except for ϕ_k and ϕ_q . These are optimized using AdamW with a batch size of 1 and a learning rate of 10^{-3} , following the value distillation objective described in Equation (19). Teacher activations for distillation are obtained by sampling with 50 denoising steps and a guidance scale of 5, using the Euler Ancestral Discrete Scheduler to integrate the reverse diffusion process.

Finetuning. In the finetuning stage, we update all parameters of the hybrid DiT, including the ϕ transformations and feed-forward MLP layers. Training uses AdamW with a batch size of 16, a learning rate of 10^{-5} , and bf16 mixed-precision. The model is trained for 1,000 iterations.

Sampling. For generating videos for VBench evaluation, we employ Wan Enhanced prompts and the following sampling configuration: 50 denoising iterations, classifier guidance scale of 6, and the UniPCMultistep noise scheduler with a flow shift of 8.

1.2. Further experimental analysis

Based on the empirical results observed in the results section, it appears that ReHyAt improves mainly on the semantic-oriented dimensions such as Object Class, Multiple Objects, Human Action, Spatial Relationship, Scene dimensions, while deteriorates on local and higher frequency categories such as imaging and aesthetic quality. We attribute this trade-off to high frequency imaging features being more sensitive to the coarser and higher-entropy attention distribution coming from linear attention, while semantic information are typically coarser in nature and less sensitive to inaccuracies from linear attention.

1.3. Trade-off analysis

We observe that our method obtains optimized hybrid models with very modest training cost and yields substantial compute and memory savings, especially at higher token counts, while largely preserving quality for moderate conversion levels (e.g. 15 or 20 blocks), as supported by benchmarks, and human studies. When converting most blocks (25 or 30 of 30), we observe some loss in temporal qual-

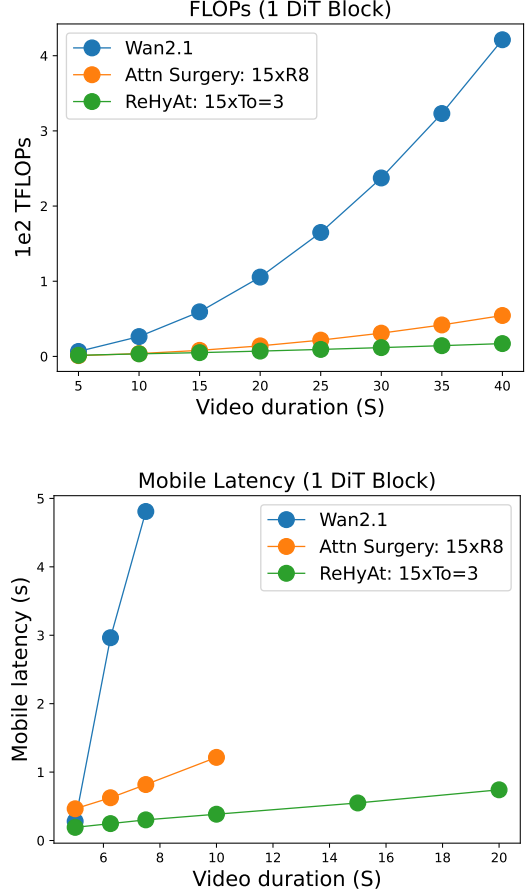


Figure 1. Compute complexity growth comparisons w.r.t. video length versus Wan2.1 flash attention and attention surgery, in FLOPs (top) and latency (bottom)

ity, reflecting the reduced expressiveness of linearized attention.

1.4. Compute complexity vs Attention Surgery

Figure 1 shows a comparison of our recurrent hybrid attention block in terms of scalability with respect to the video length versus attention surgery hybrid and original Wan2.1 flash attention blocks.

1.5. Comparison with token merging/pruning

In Tab. 1, we compare VBench scores as reported in USV (Wu et.al 25) for two token merging methods, vanilla ToMe (Boyla et.al 22) and USV dynamic sparsity allocation, both similarly applied on Wan2.1 1.3B @480p resolution.

Table 1. Comparison with Token merging methods

Model (Wan2.1 1.3B)+	ReHyAt	ToMe	USV
VBench Total	83.8	77.1	80.7

1.6. Two-stage algorithm

We further evaluated two hybrid-attention configurations with and without the layer-wise distillation stage. As shown in Tab. 2, distillation provides a consistent but modest improvement, indicating that a one-stage training scheme already yields strong results. This analysis can provide the reader with support in deciding between (slightly) better and faster convergence or implementation simplicity.

Table 2. VBench Score; impact of pretraining is consistent but not large

Stage1	ReHyAt- $25 \times T_c=1$	ReHyAt- $25 \times T_c=2$
×	80.7	81.7
✓	81.0	82.1

1.7. Qualitative Samples

Figures 3–19 present uniformly spaced frames from videos generated by the original Wan2.1 1.3B model and several variants of our recurrent hybrid attention models ($15 \times T_c=5$, $15 \times T_c=3$, and $20 \times T_c=3$) across 18 prompts at the original resolution of 480×832 . Full video sequences corresponding to these frames are included in the supplementary materials.

1.8. Detailed VBench Comparison

Figure 2 compares a selected subset of our hybrid models against Wan2.1 1.3B across all VBench dimensions, evaluated on the full benchmark set at the original resolution (480×832).

1.9. Detailed VBench-2.0 Comparison

Tables 3–5 report fine-grained results on the recent VBench-2.0 benchmark at 480×832 resolution. We compare two ReHyAt variants ($15 \times T_c=3$ and $15 \times T_c=5$) against Wan2.1 1.3B and attention surgery ($15 \times R2$). Both hybrid variants perform on par with Wan2.1 1.3B in terms of the overall Total score.

1.10. Use of Large Language Models

We used Microsoft Copilot (a large language model) exclusively to improve clarity and readability. All technical content, experimental design, and conclusions are entirely our own.

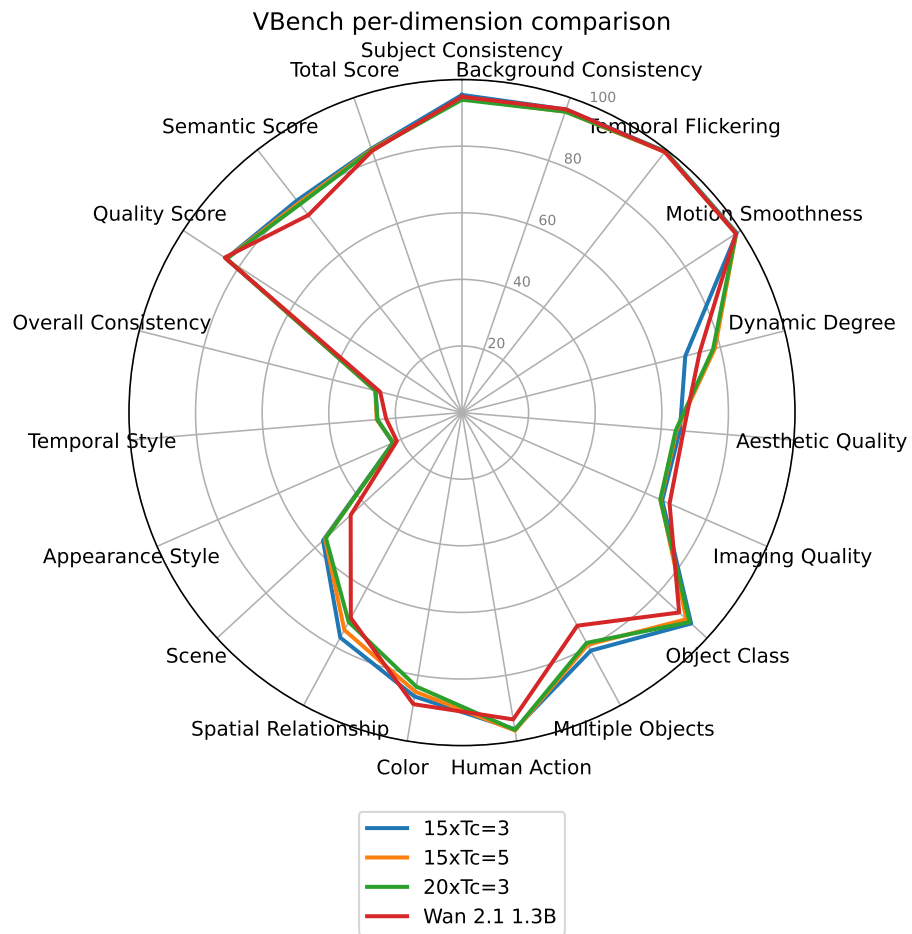


Figure 2. Radar plot comparing a subset of our hybrid models with the original Wan 1.3B model on the full VBBench set and 480×832 resolution

Method	Human Identity	Dynamic Spatial Relationship	Complex Landscape	Instance Preservation	Multi-View Consistency	Human Clothes	Dynamic Attribute	Complex Plot
Wan2.1 1.3B*	63.5	25.1	16.4	86.0	9.6	97.9	49.1	11.3
Attention Surgery (15×R2)	62.7	25.1	18.4	84.8	7.1	97.1	44.0	13.2
RehHyAt 15× $T_c=3$	64.7	28.5	14.7	78.4	12.1	98.1	22.0	12.7
RehHyAt 15× $T_c=5$	61.6	28.0	16.7	83.6	10.6	94.2	28.6	15.6

Table 3. Full VBench-2.0 results (part 1/3).

Method	Mechanics	Human Anatomy	Composition	Human Interaction	Motion Rationality	Material	Diversity	Motion Order Understanding
Wan2.1 1.3B*	72.4	80.6	48.4	71.7	40.8	69.4	49.1	32.0
Attention Surgery (15×R2)	66.4	77.0	46.4	70.3	41.4	67.3	48.5	33.7
RehHyAt 15× $T_c=3$	63.7	83.0	46.4	75.0	47.1	69.6	63.8	37.0
RehHyAt 15× $T_c=5$	64.7	83.6	51.0	72.3	44.8	67.8	60.4	34.3

Table 4. Full VBench-2.0 results (part 2/3).

Method	Camera Motion	Thermotics	Creativity Score	Commonsense Score	Controllability Score	Human Fidelity Score	Physics Score	Total Score
Wan2.1 1.3B*	32.1	61.7	48.7	63.4	34.0	80.7	53.3	56.0
Attention Surgery (15×R2)	29.0	70.5	47.5	63.1	33.4	79.0	52.8	55.1
RehHyAt 15× $T_c=3$	25.9	54.6	55.1	62.7	30.8	81.9	50.0	56.1
RehHyAt 15× $T_c=5$	29.0	55.7	55.7	64.2	31.9	79.8	49.7	56.3

Table 5. Full VBench-2.0 results (part 3/3).

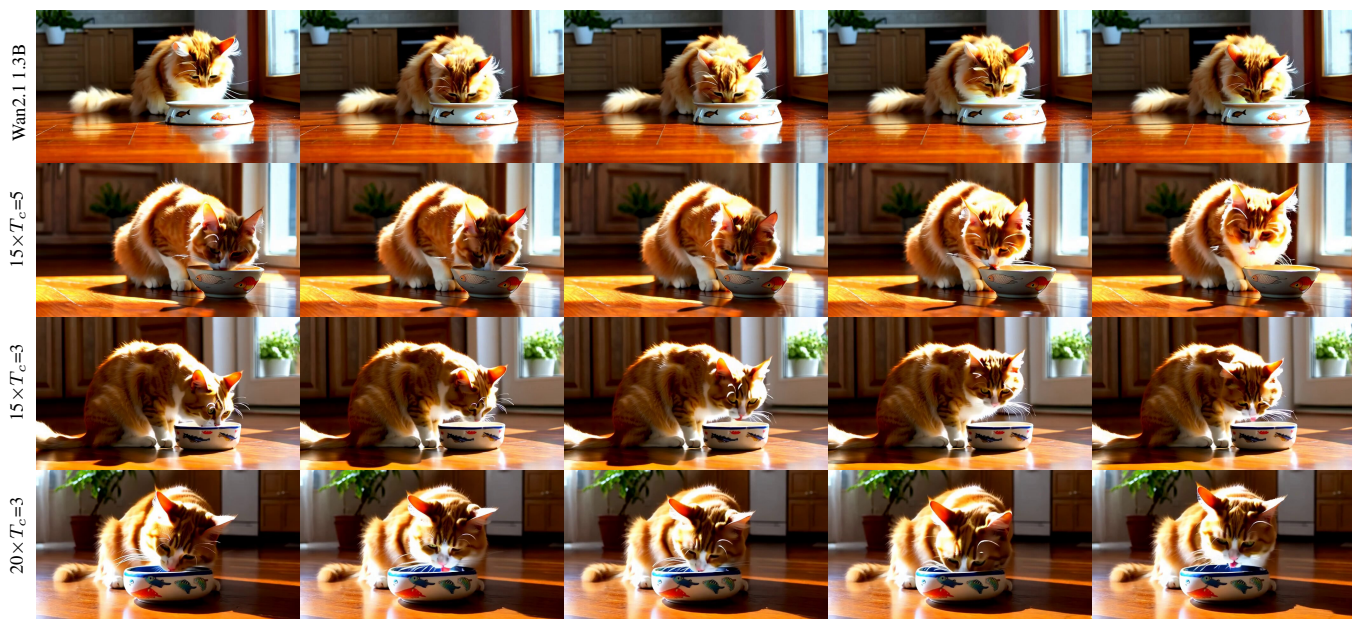


Figure 3. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A cat eating food out of a bowl*



Figure 4. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *a person playing guitar*



Figure 5. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A cute fluffy panda eating Chinese food in a restaurant*



Figure 6. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A cute happy Corgi playing in park, sunset, with an intense shaking effect*



Figure 7. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *a dog running happily*



Figure 8. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A fat rabbit wearing a purple robe walking through a fantasy landscape.*



Figure 9. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A person is crying*



Figure 10. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *a cow bending down to drink water from a river*



Figure 11. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A bigfoot walking in the snowstorm*.



Figure 12. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *a bear sniffing the air for scents of food*

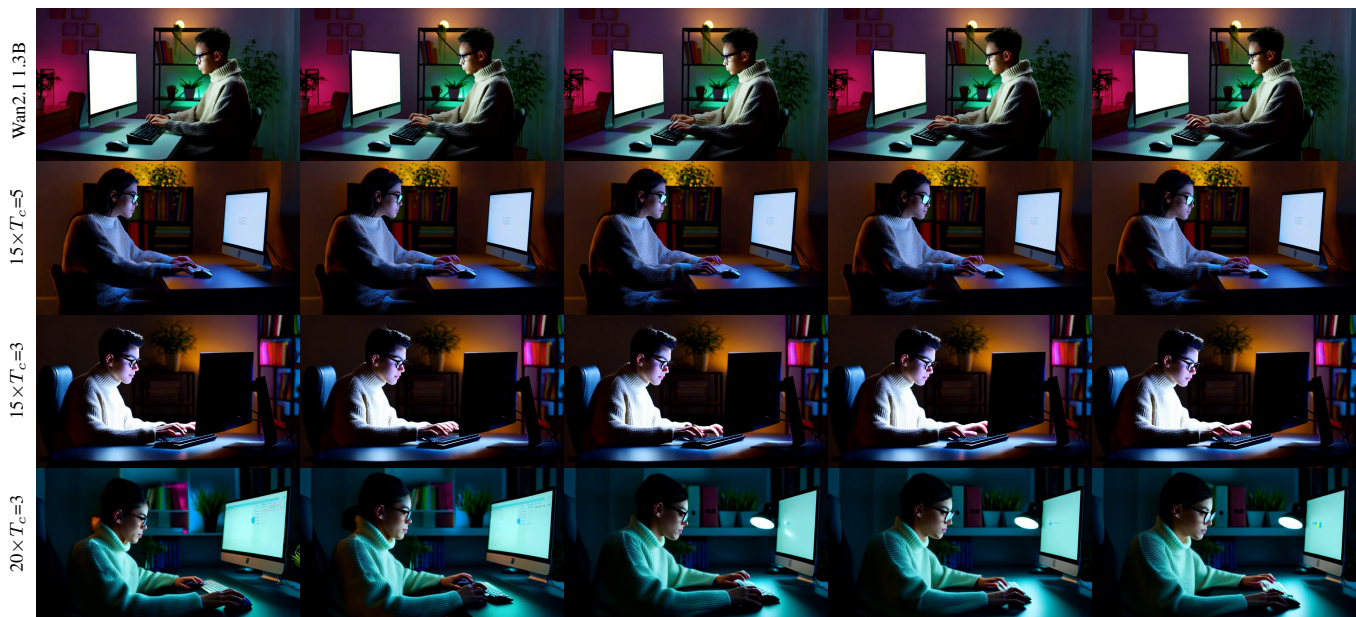


Figure 13. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A person is using computer*



Figure 14. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *a sheep taking a peaceful walk*

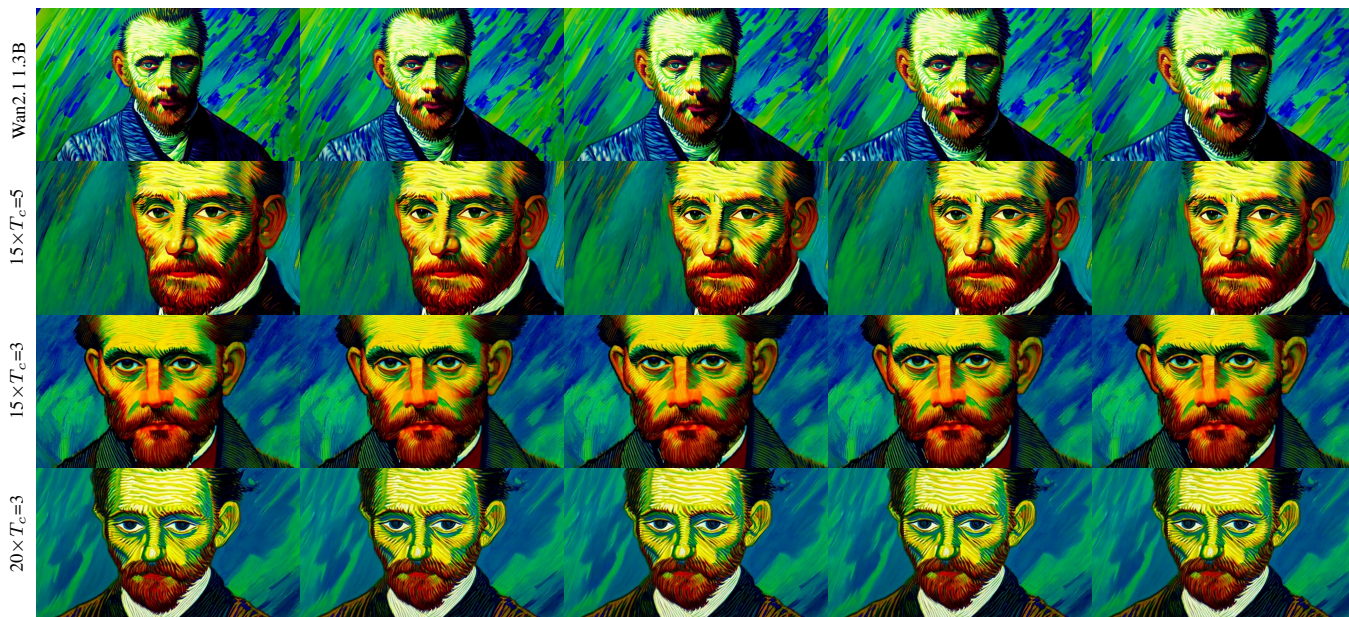


Figure 15. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *Cinematic shot of Van Gogh's selfie, Van Gogh style*



Figure 16. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *happy dog wearing a yellow turtleneck, studio, portrait, facing camera, dark background*



Figure 17. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *this is how I do makeup in the morning*.

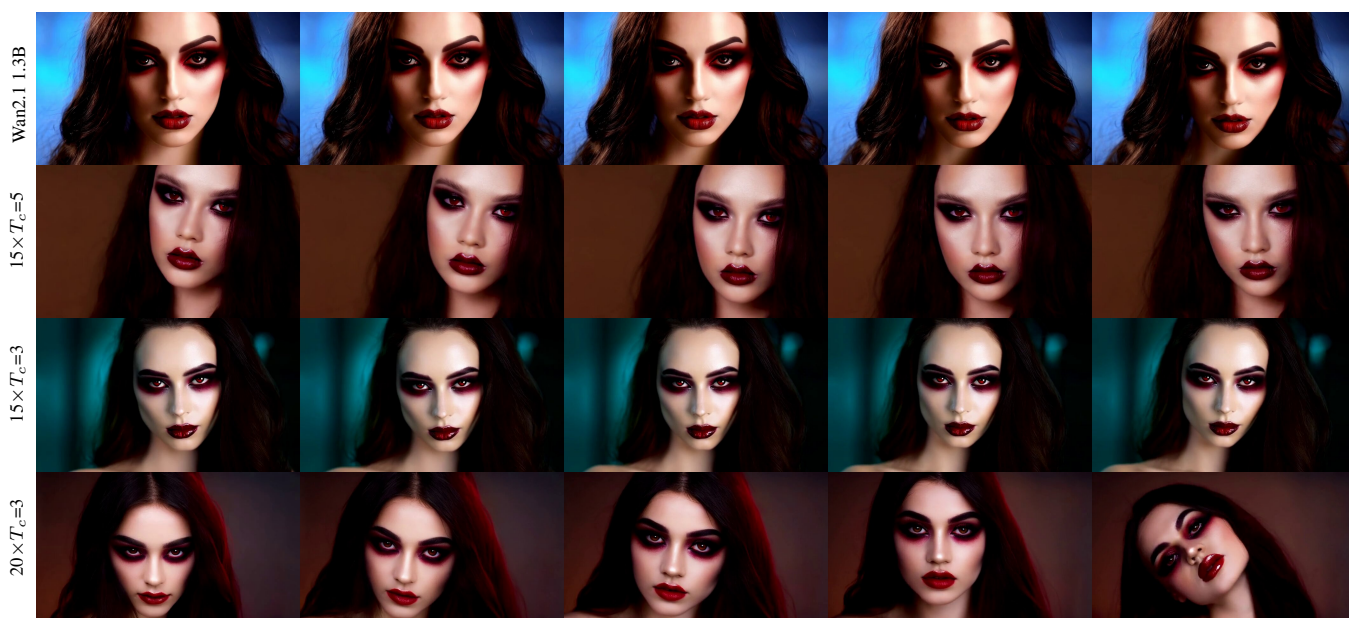


Figure 18. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *Vampire makeup face of beautiful girl, red contact lenses*.

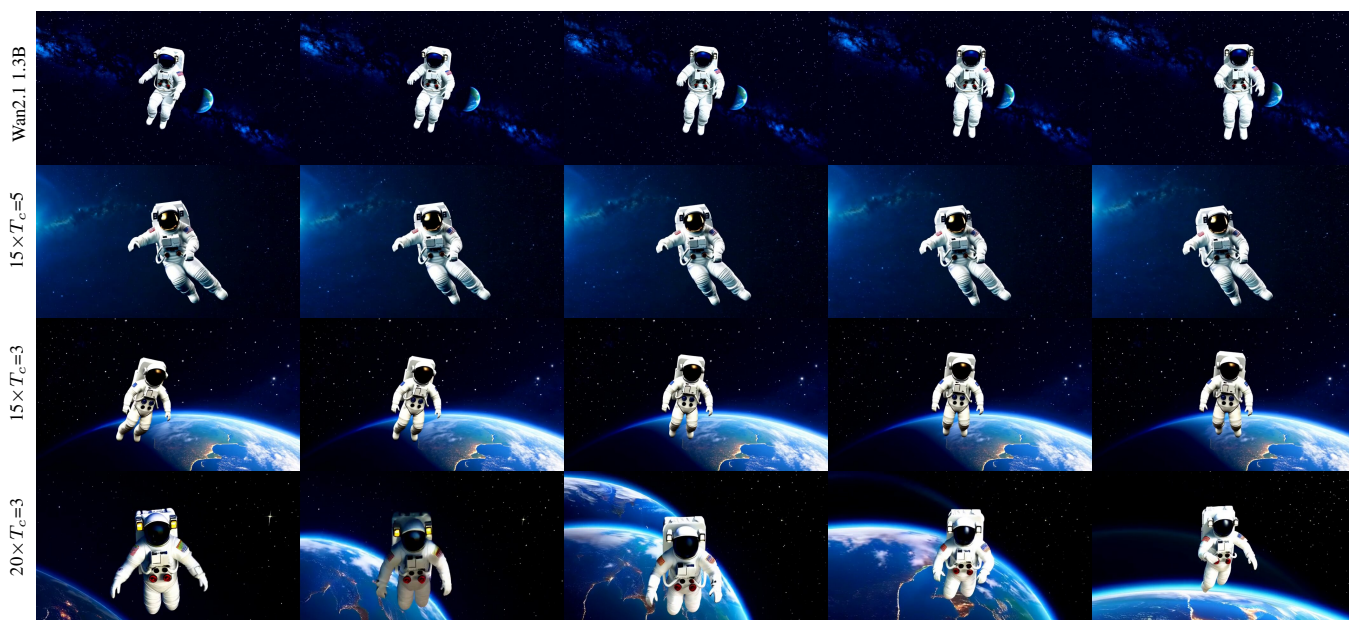


Figure 19. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *An astronaut flying in space, featuring a steady and smooth perspective*