Attention Surgery: An Efficient Recipe to Linearize Your Video Diffusion Transformer

Supplementary Material

1. Appendix

1.1. Training Details and Hyperparameters

Except for the ablation studies we characterize ϕ with a 2-layer MLP and a polynomial degree of 2, and make two separate transformations for keys and queries (ϕ_k and ϕ_q) per hybrid block.

Pretraining distillation stage. We train each block independently and all the parameters are frozen except for the ϕ_k and ϕ_q, for which we use the AdamW optimizer, batch size of 1 and a learning rate of 1e-3, with the value distillation objective, as detailed in equation (7) to train. To extract teacher activations for distillation, we sample using 50 denoising steps with a guidance scale of 5, employing the Euler Ancestral Discrete Scheduler to integrate the reverse diffusion process.

Finetuning. Within the finetuning process, we finetune all parameters of the hybrid DiT, including the ϕ's, the feed-forward MLP, etc., with a batch size of 16, AdamW optimizer and a learning rate of 1e-5 and bf16 mixed precision training. The model is trained for only 1000 iterations.

Sampling. For sampling videos for VBench, we use the Wan Enhanced prompts and the following sampling parameters: We sample all the variants with 50 denoising iterations, classifier guidance scale of 6 and UniPCMultistep noise Scheduler with flow shift of 8.

1.2. GPU latencies

While uniform softmax-token subsampling introduces some irregular memory access, in our design the index/gather occurs only once, is independent of the query tokens, and the resulting tensors are made contiguous for subsequent operations. The linear branch runs entirely as dense GEMMs, and the softmax branch operates on a reduced contiguous sequence (not a masked sparse pattern). Thus, the removal of the quadratic softmax term outweighs the gather overhead in practice. Table 1 represents a block level latency profiling on an A100.

Table 1. GPU latency (ms) profiling for various video resolutions.

Attn. architecture	720×1280×81	480×832×81	320×480×81
Softmax (Flash Att.)	232.0	58.3	16.9
Ours Hybrid R8	**157.4**	**47.5**	**15.8**

1.3. Uniform vs local softmax tokens

Our design choice of uniform-temporal subsampling decouples the k/v selection from query indices and, more importantly, avoids the temporal drift and unsmoothness we observe with local-temporal softmax. In Tab. 2, we compare our method (25×R4) to a computationally-matched rearranged hybrid variant with local-temporal softmax, measuring temporal qualities such as mean warping error and temporal jitter on VBench samples.

Table 2. Comparing temporal quality of various attention types

	Mean Temporal Jitter ↓	Mean Warping Error ↓
Tempo-Local	1.06	0.091
Tempo-Uniform (ours)	**0.77**	**0.064**

1.4. Application to a second/larger model

As an additional evidence for the effectiveness and generalizability of our method, we integrated Attention surgery to Wan**2.2 5B**, a larger and newer model compared to our main Wan2.1 1.3B baseline. Tab. 3 shows the results.

Table 3. Wan2.2 **5B** model @480×832×81 resolution

Model	Wan2.2 5B	Ours 15×R2	Ours 15×R4
VBench Total	82.23	82.23	81.51

1.5. Vs. token reduction

In Tab. 4, we present a comparison with ToMe (Boyla et al. 23) and USV (Wu et. al, 25) both reporting lower VBench scores applying on Wan2.1 1.3B @480p, similarly. Notably, token merging introduces content-dependent sparsification, whereas our approach is a training-light retrofit that preserves full global coverage and is complementary to merging-based accelerations.

Table 4. Comparison with Token merging methods

Model (Wan2.1 1.3B)+	Ours	ToMe	USV
VBench Total	**83.1**	77.1	80.7

1.6. Qualitative samples

Figures 2 to 19 show uniformly spaced frames from videos generated by the original Wan2.1 1.3B and different variations of our hybrid attention models (15×R2, 15×R4,

15×R8, 20×R4, and 20×R8), for 18 different prompts on the original 480×832 resolution. All the videos corresponding the demonstrated frames, are available as video files in the attached supplementary materials.

1.7. Memory Read/Write on Mobile

Table 5 shows the the memory read/write values for one DiT block in various methods in GB, as measured with the QNN runtime on a Snapdragon8-Gen4 SoC.

1.8. Detailed VBench Comparison

Figure 1 shows a selected subset of our hybrid models compared to the original Wan2.1 1.3B model, on all of the comparison dimensions. The experiment is with the full VBench set and at the original 480×832 resolution.

1.9. Detailed VBench-2.0 Comparison

Tabs. 6 to 8 demonstrate fine-grained results on the recent VBench-2.0 benchmark at original resolution of 480×832. We generated videos with the original Wan2.1 1.3B model and our 15×R2 modification using the same sampler hyperparameters. We observe that our hybrid model experiences an insignificant drop in performance as measured by Total score.

1.10. Use of Large Language Models

We used Microsoft Copilot (a large language model) to aid in polishing the writing of this submission. The model was employed solely for improving clarity and readability; all ideas, technical content, and conclusions are our own.

Attention Block	Number of frames - Memory Read/Write (GB)							
	81		101		121		141	
	W	R	W	R	W	R	W	R
Softmax Flash Attention	5.1	6.0	12.9	16.4	22.7	53.6	OOM	OOM
HedgeHog Linear Attention	5.7	8.1	7.0	10.1	6.9	11.3	8.0	13.2
Attention Surgery - R8	6.3	10.1	5.2	10.9	6.4	13.2	7.8	35.2

Table 5. Comparison of total memory read/write for Wan2.1 DiT Blocks with various attention mechanisms on Snapdragon8-Gen4

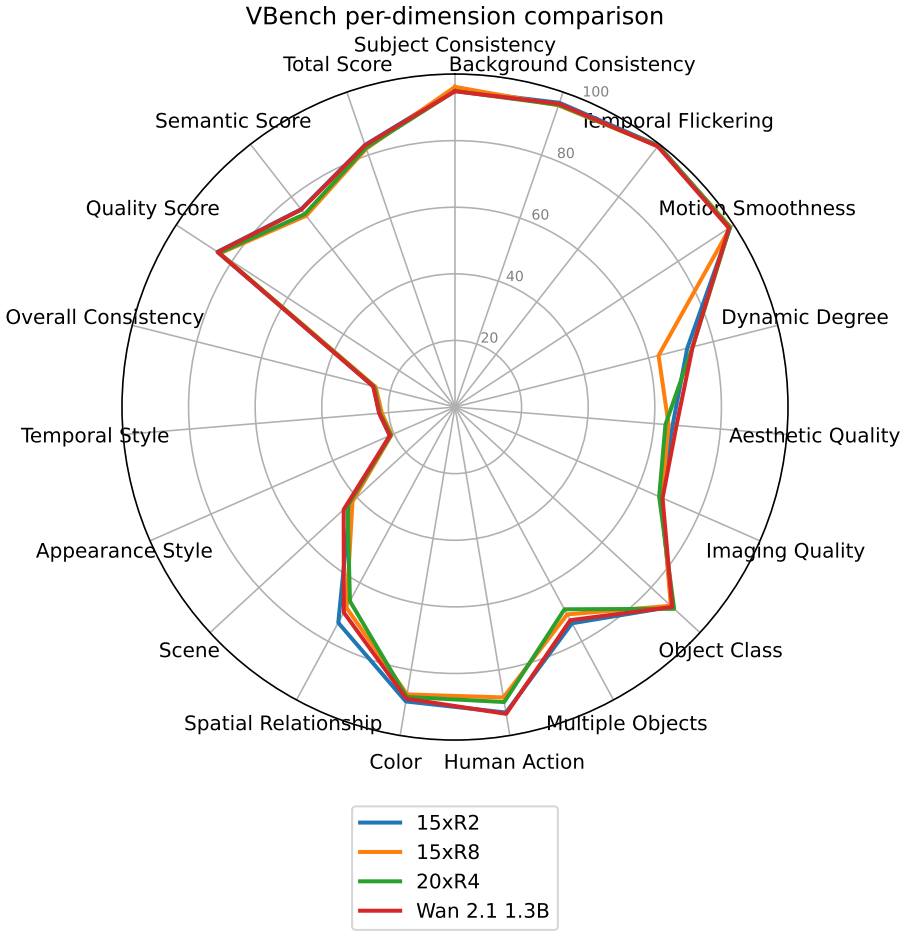

Figure 1. Radar plot comparing a subset of our hybrid models with the original Wan 1.3B model on the full VBench set and 480×832 resolution

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

Table 6. 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

Table 7. 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

Table 8. Full VBench-2.0 results (part 3/3).

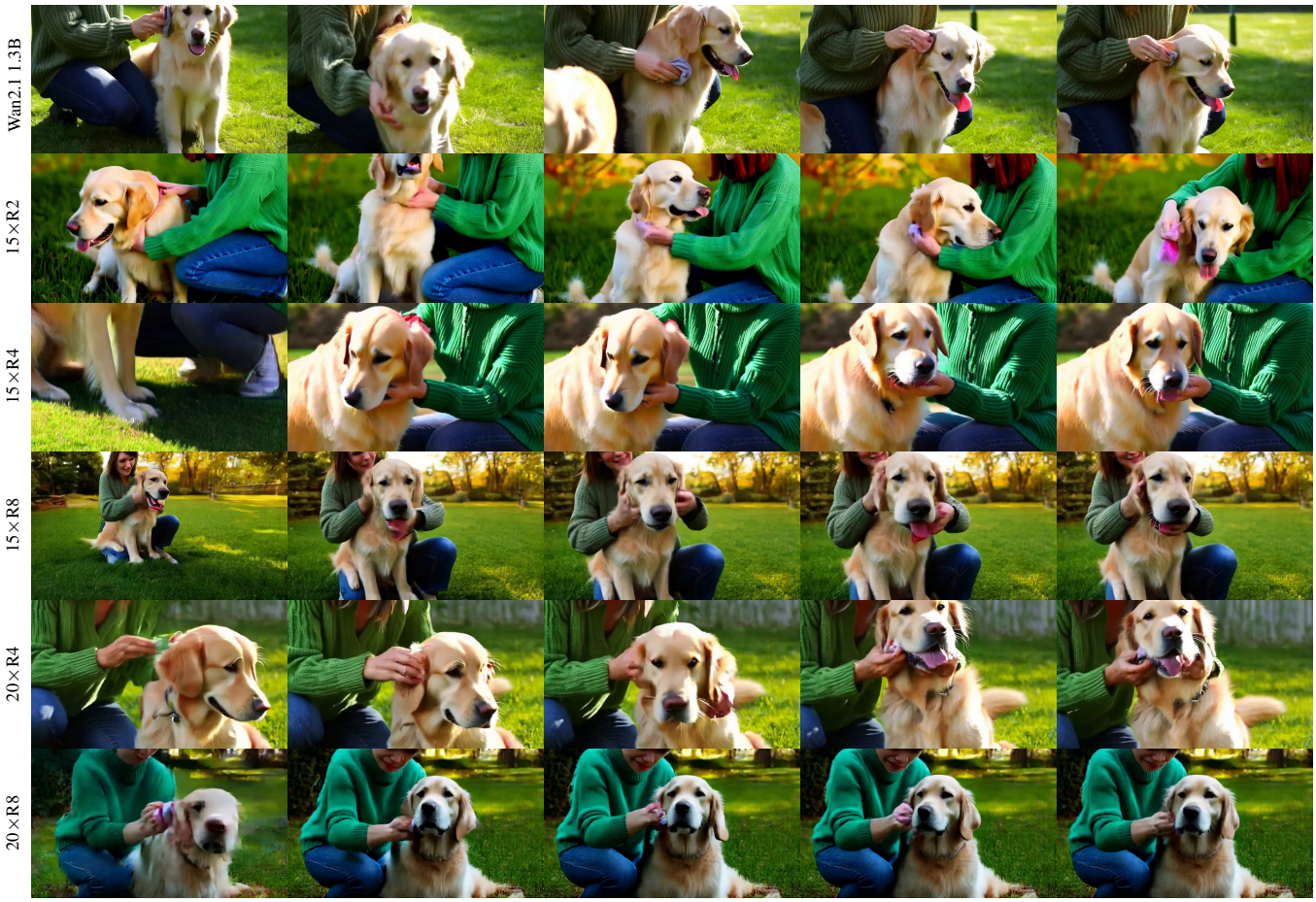

Figure 2. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A person is grooming dog*

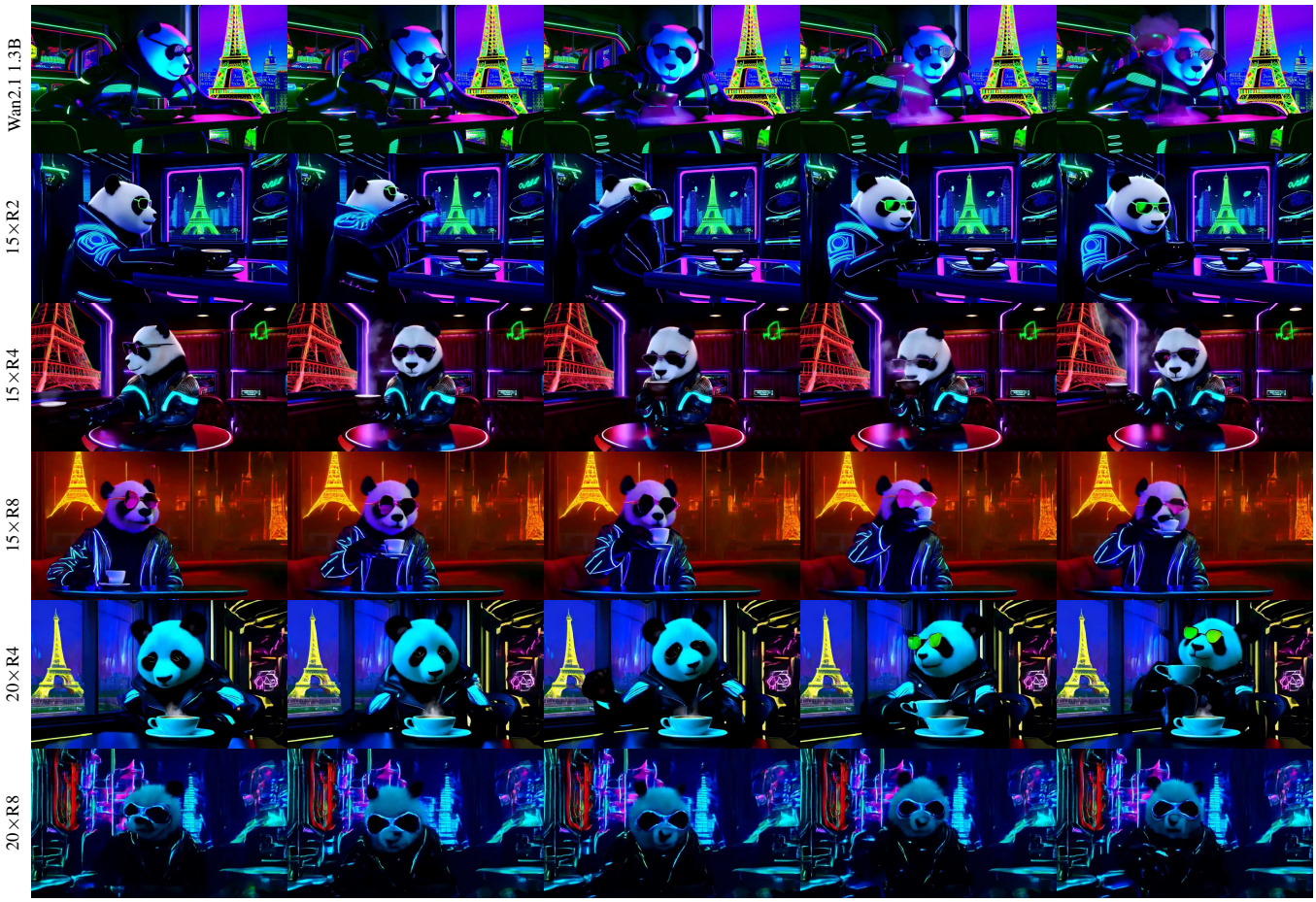

Figure 4. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A panda drinking coffee in a cafe in Paris, in cyberpunk style*

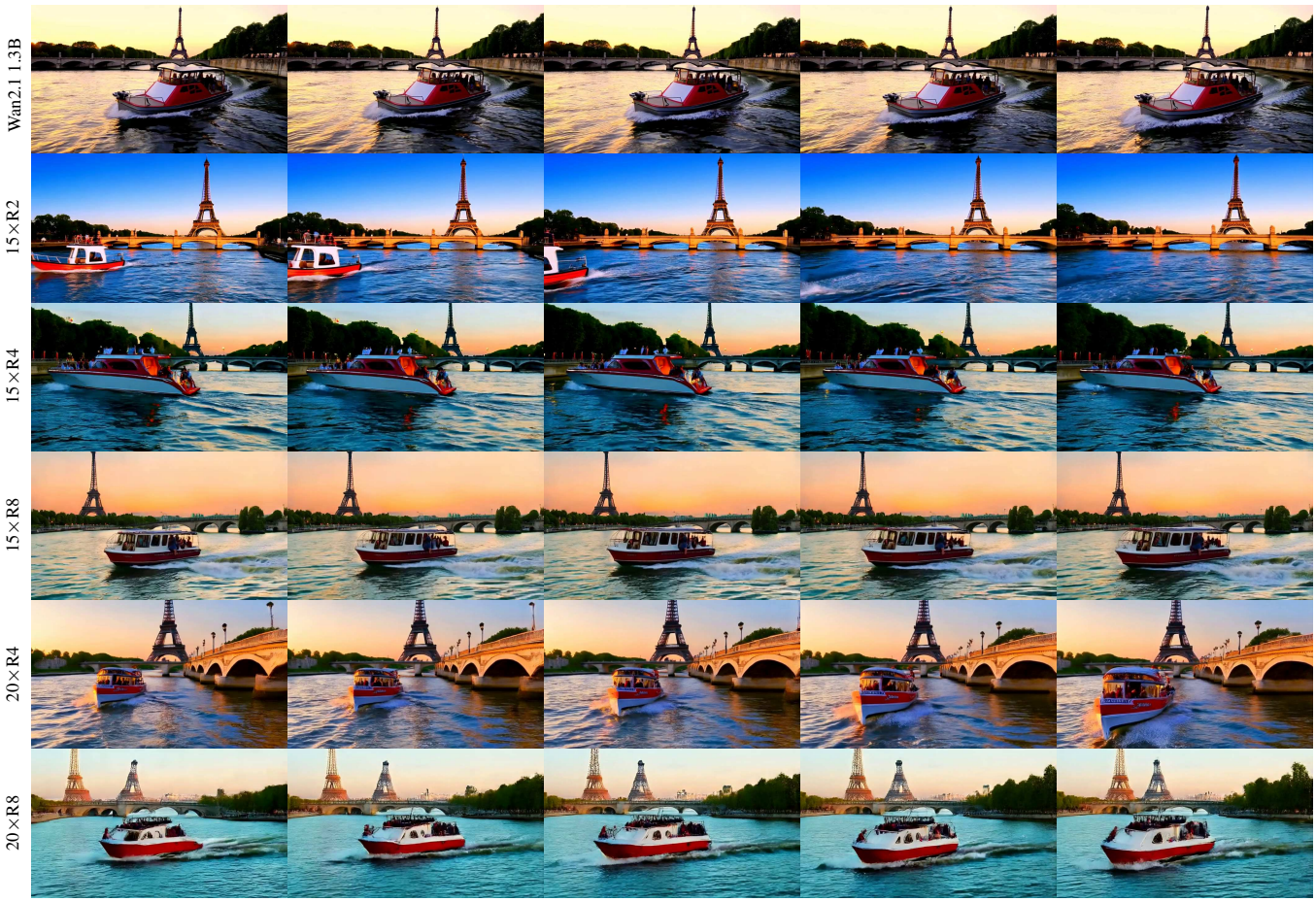

Figure 5. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A boat sailing leisurely along the Seine River with the Eiffel Tower in background*

Figure 6. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *a black cat*

Figure 7. 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 8. 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*

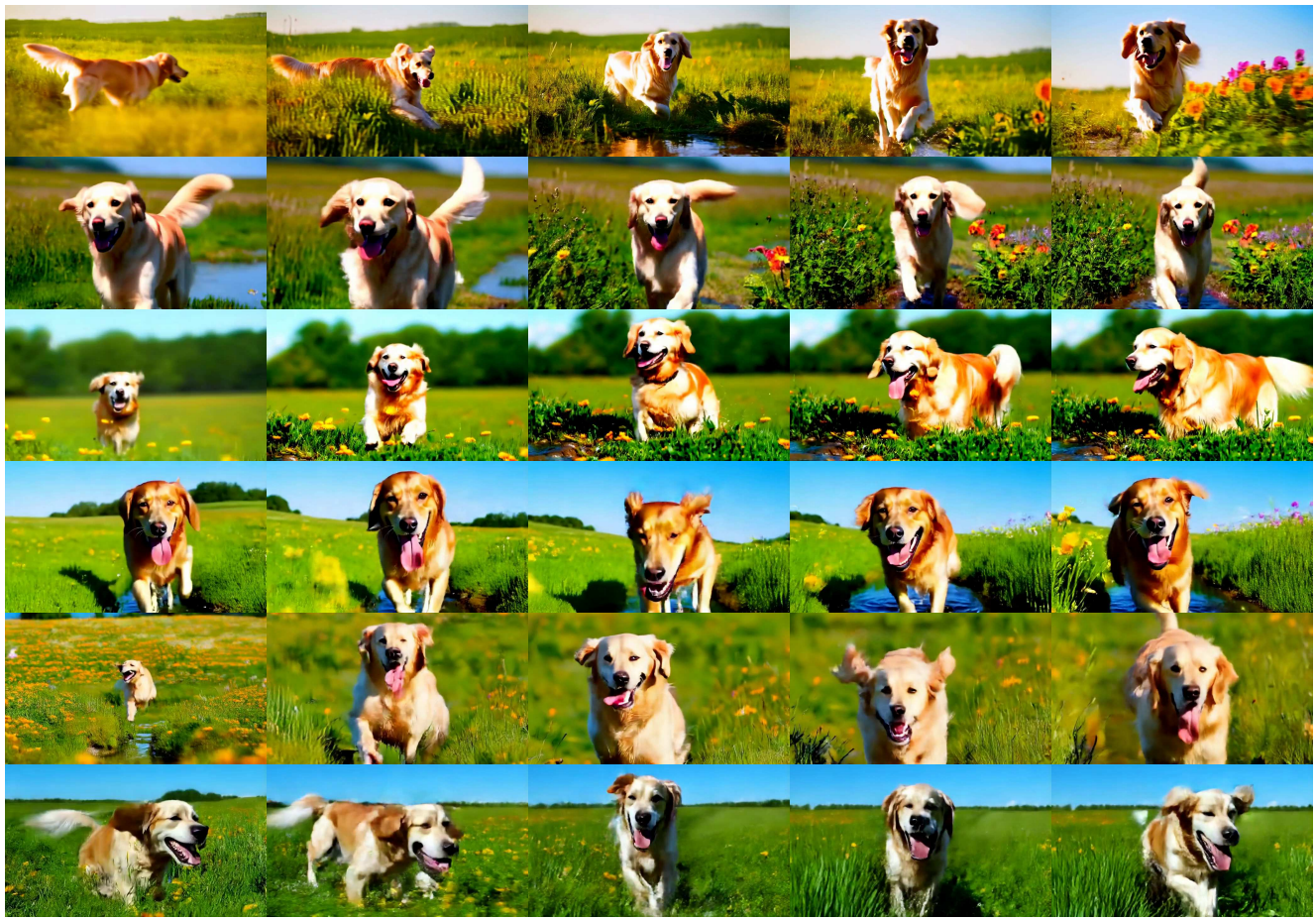

Figure 9. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *a dog running happily*

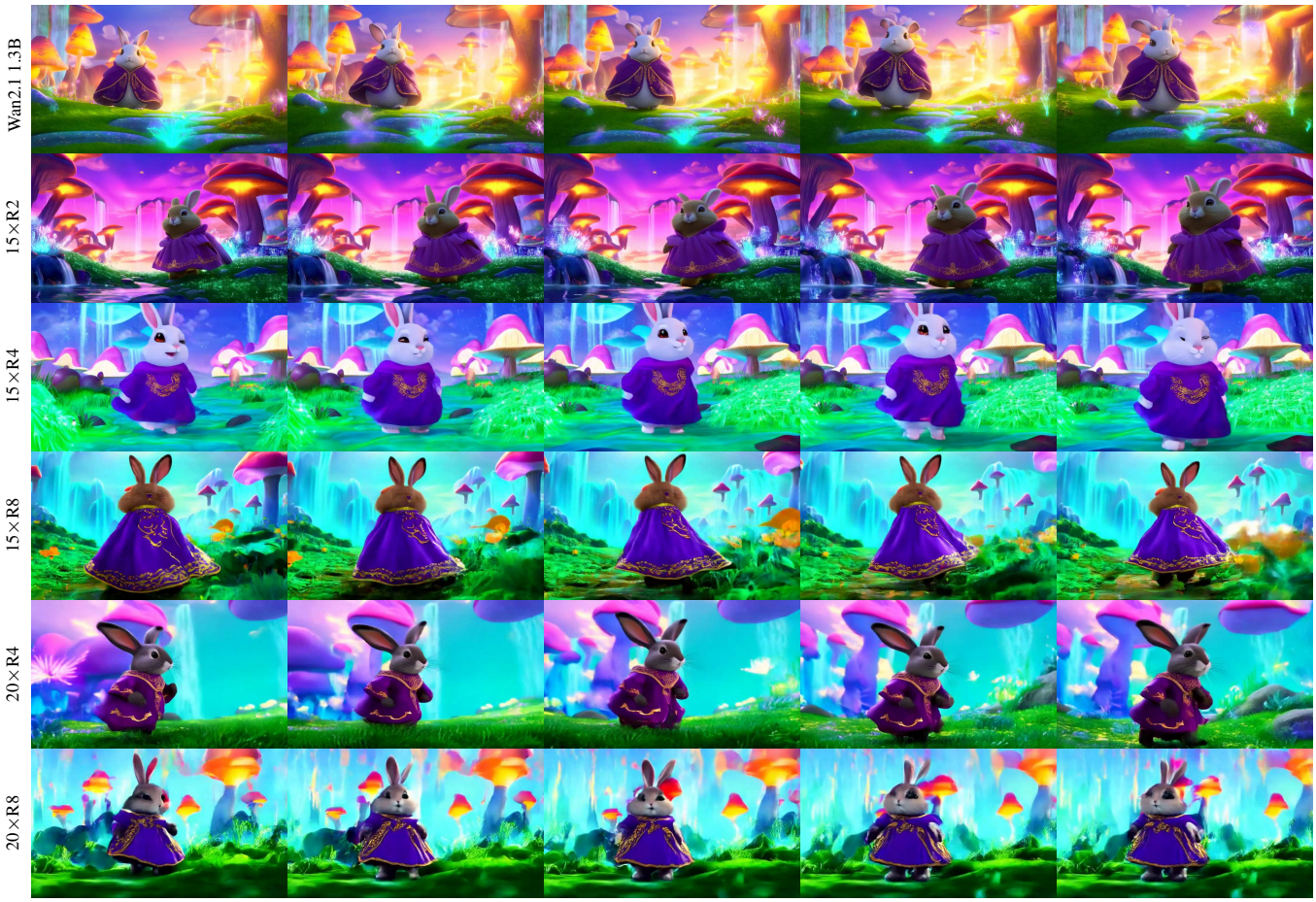

Figure 10. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A fat rabbit wearing a purple robe walking through a fantasy landscape.*

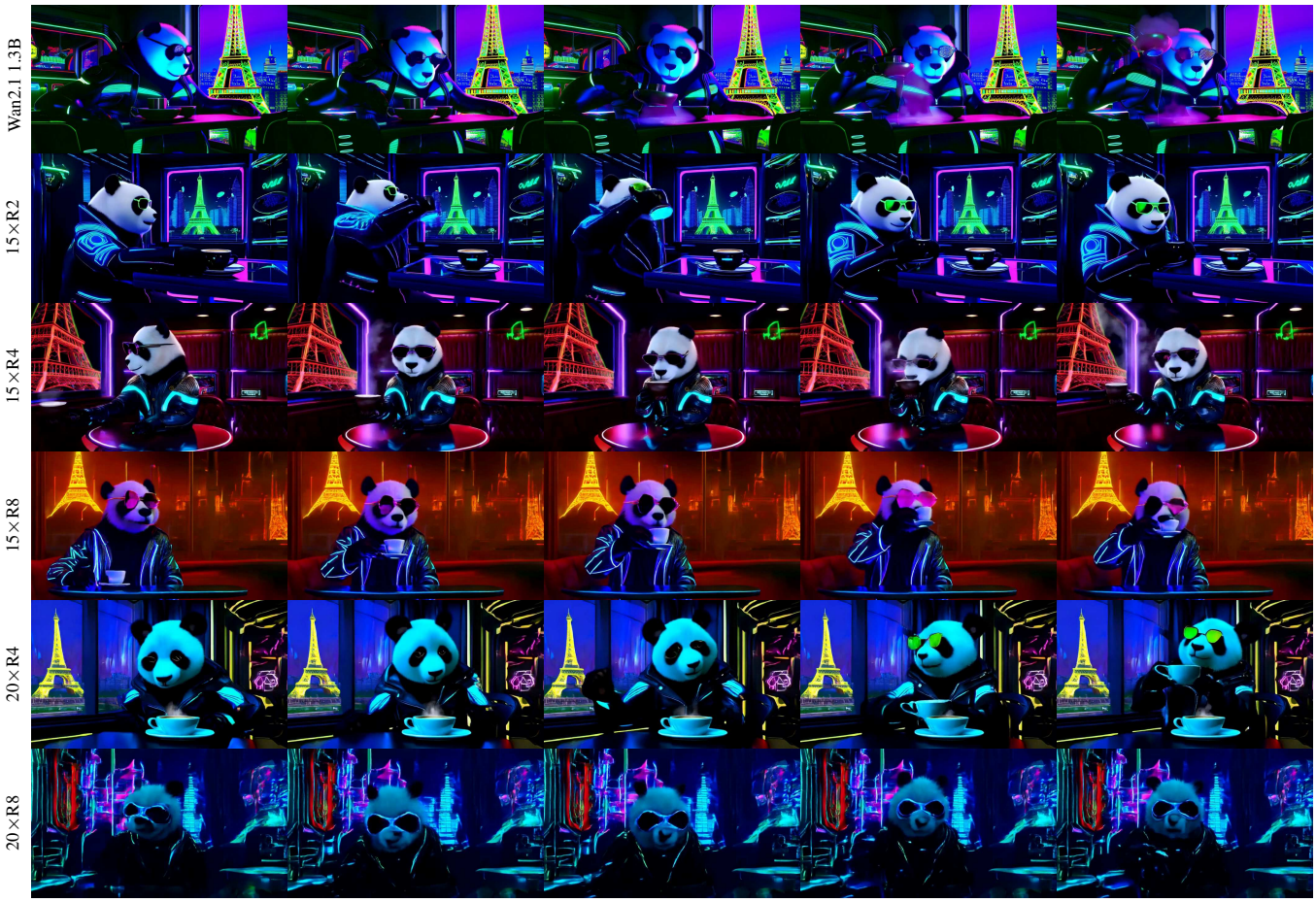

Figure 11. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A panda drinking coffee in a cafe in Paris, in cyberpunk style*

Figure 12. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *a teddy bear is swimming in the ocean*

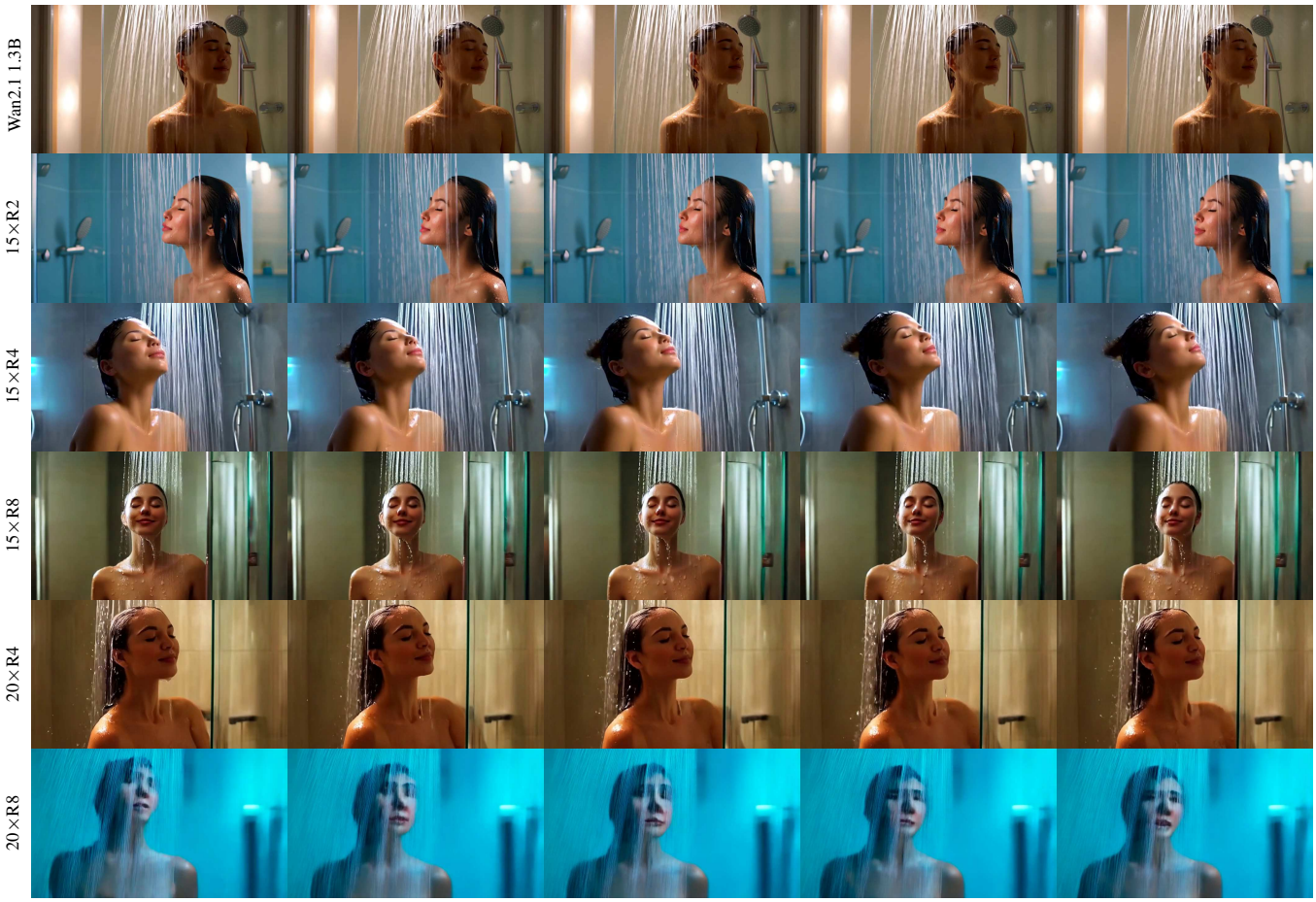

Figure 13. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A person is taking a shower*

Figure 14. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *A person is using computer*

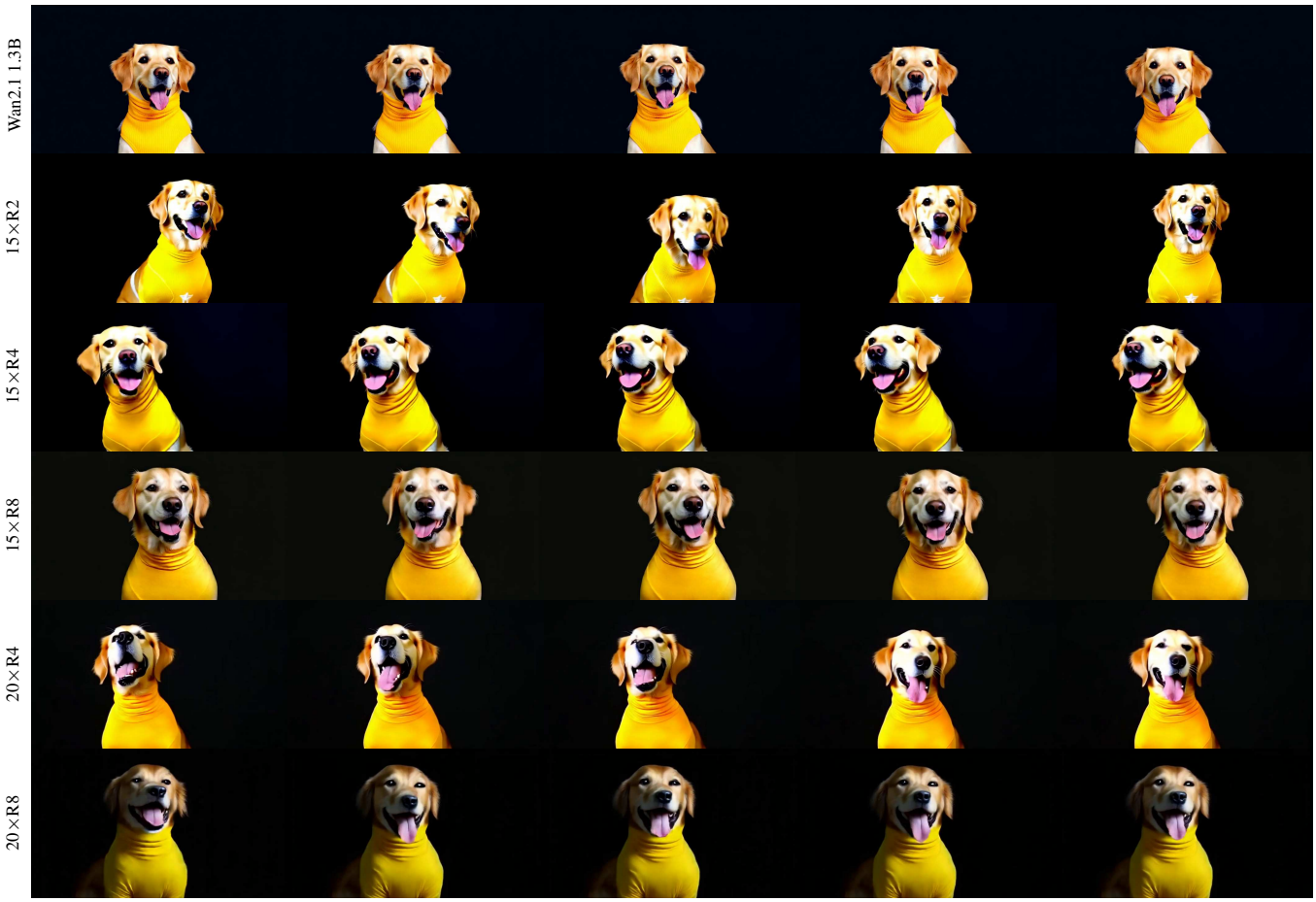

Figure 15. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *happy dog wearing a yellow turtleneck, studio, portrait, facing camera, dark background*

Figure 16. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *Iron Man flying in the sky*

Figure 17. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *raceway*

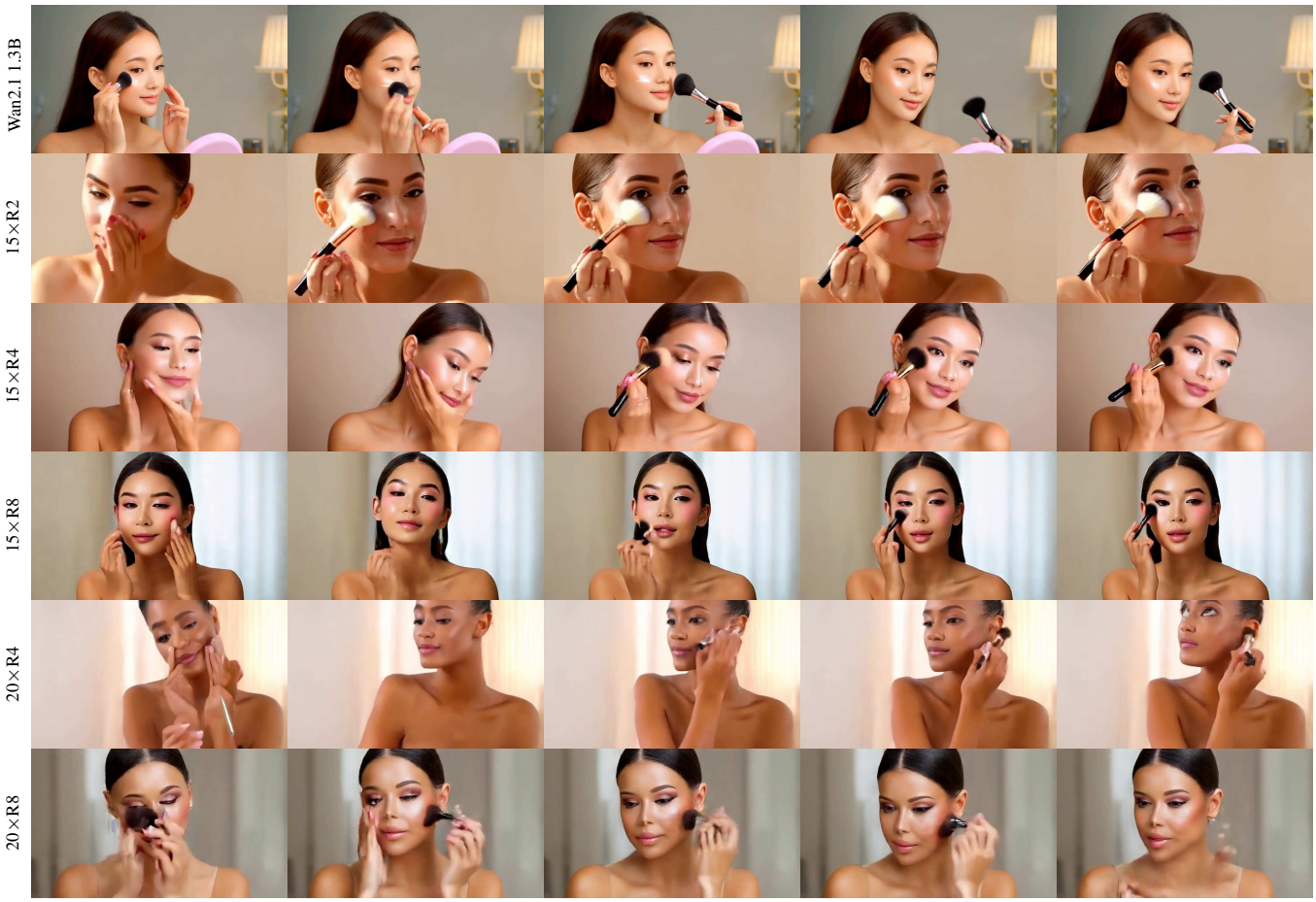

Figure 18. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *this is how I do makeup in the morning.*

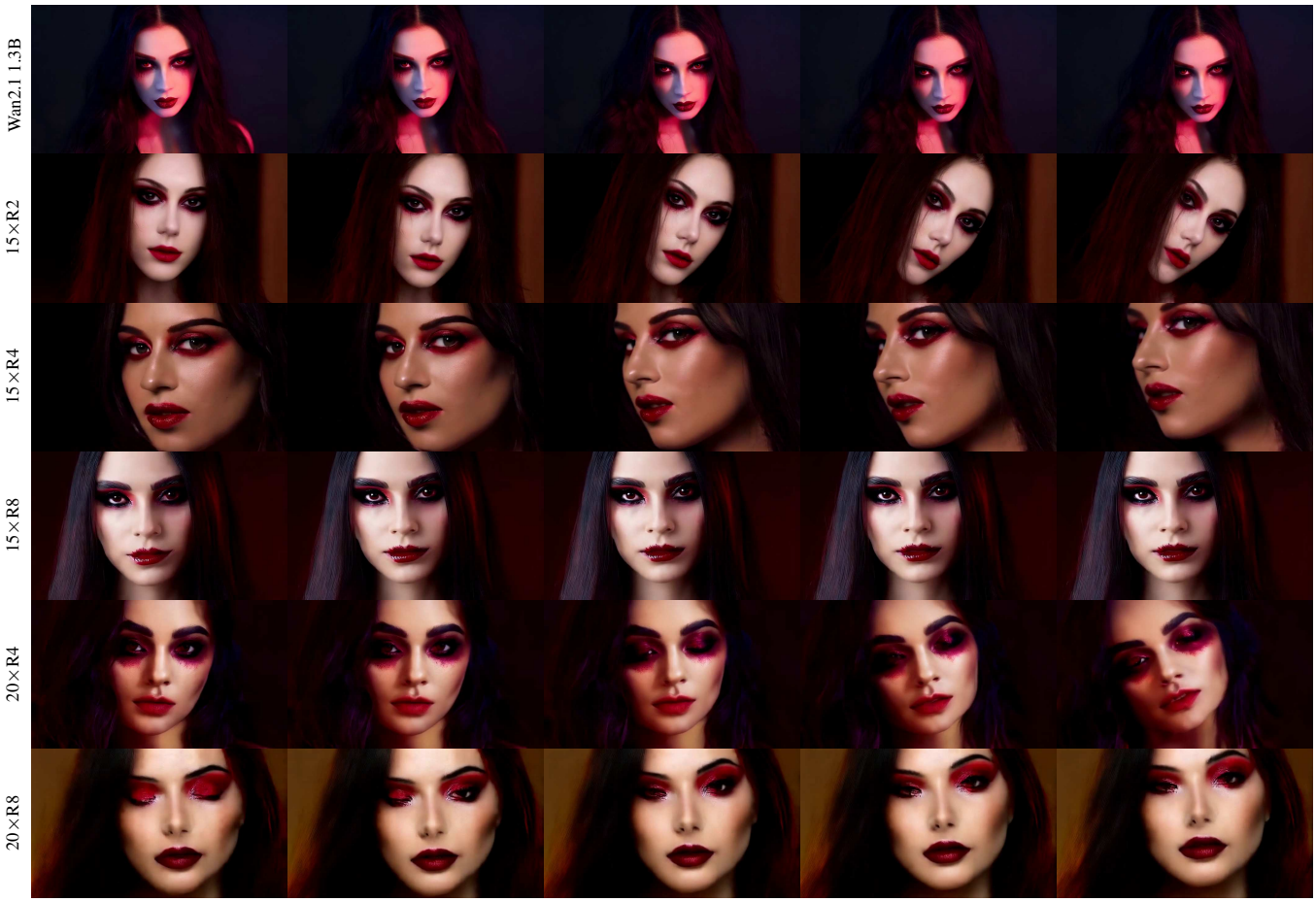

Figure 19. Qualitative videos comparing original Wan2.1 1.3B model to our various hybrid variations for input prompt *Vampire makeup face of beautiful girl, red contact lenses.*