

Block Cascading: Training Free Acceleration of Block-Causal Video Models

Supplementary Material

Multi GPU Inference

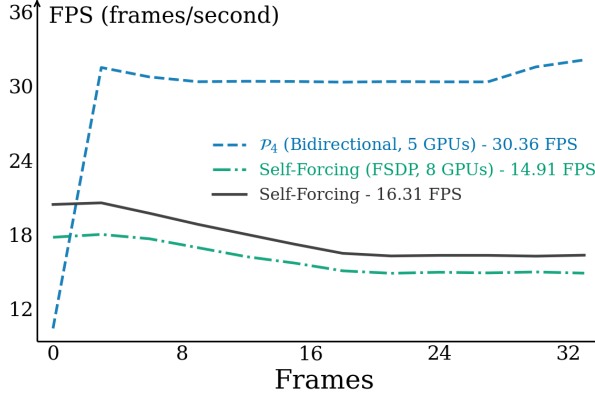


Figure S1. **Multi-GPU Inference:** Forcing multi-GPU inference with FSDP over Self-Forcing [3] block-causal pipeline

Table S1. Parallelising baseline SF across more GPUs

SF Parallelism	2 GPUs	3 GPUs	4 GPUs	5 GPUs	6 GPUs
Ring Attention	1.66	1.67	1.56	1.60	1.49
Ulysses	18.15	18.12	18.80	–	18.77
FSDP	15.30	15.09	15.24	14.96	14.94

S1. Multi-GPU block-causal generation

To demonstrate the lack of multi-GPU support for block-causal inference, we split model parameters across GPUs (FSDP) as a form of single-prompt parallelism possible, using block-causal baseline Self-Forcing [3]. In here, the model is evenly sharded (FSDP) across 8 GPUs (*i.e.* each GPU sees $\frac{1}{8}$ th of the model). We find that this parallelism does not improve FPS rather burdens the pipeline with too many synchronizations and slows down inference further. We summarise our results in Fig. S1. We attempt other parallelisation methods (Tab. S1) like Ulysses parallelization, which offers minimal speedup on Wan 1.3B (12 heads) because of communication overhead: at (max) 6-way parallelism, it gets ~ 18.8 FPS vs. our 30.3 FPS. Ring Attention parallelizes sequence (optimizing memory) but is slower

S2. Qualitative Analysis

We conduct additional qualitative analysis with full bidirectional attention in Fig. S2, demonstrating that (noisy) bidirectional inference is still possible with trained block-causal pipelines. We also include additional qualitative analysis comparing our Block Cascading pipeline with baselines in Fig. S6 and Fig. S7.



Figure S2. **Bidirectional Ablation:** Fully bidirectional inference with block-causal trained network Self-Forcing [3]

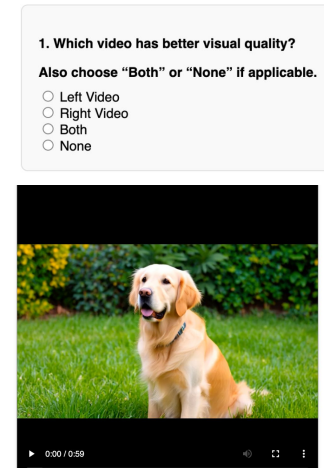


Figure S3. **User study:** Demo visual of the user study environment visible to AMT workers

S3. User study

We conduct our user study (Sec. 4.3) with 44 unique workers on Amazon Mechanical Turk (AMT) obtaining a total of 7551 votes across all comparisons (ablations and baseline comparisons), of which we retain 6149 votes. We flag workers who spend less time in analysing choices (*e.g.* < 30 s) or have a strong position bias (*e.g.* always ‘left’) with negative trust scores. Trust scores of workers go up when they agree with each other over choices. We provide an example of what the worker sees during the user study in Fig. S3. For interactive video analysis, we condense the input prompt (with Claude Sonnet 4.5 [2]) to make it shorter and improve comprehension.

S4. KV-Recaching (Contd.)

In addition to the discussion on KV-recaching in Sec. 3.5, we include a study where we analyse qualitatively the effect of reducing number of recache tokens during KV-recaching

to simulate a low-latency recaching environment. Reducing the number of tokens to re-cache prevents their computation with the new prompt thereby reducing latency. We find in Fig. S4 that with only one block being recached, quality degrades during context switching, often leading to sudden movements and blocky motion. These are general issues with KV recaching which get worse when number of re-cache frames are reduced to 3 (*i.e.* 1 block).

S5. Additional Quantitative Analysis

We include VBench [4] scores from all categories for competitors in Tab. S2, the summary of which is available in Tab. 1. We also report First-Frame Latency (FFL) in Tab. S4. Latency drops after cascade construction, steady at $\sim 0.4\text{s/block}$. \mathcal{P}_2 and \mathcal{P}_3 have FFLs 0.68 and 0.69s. SF’s latency increases from KV accumulation. We report Quality Drift and VBench-Long metrics on 30s videos. Additional quantitative analysis of cascading type and window sizes (on Self-Forcing) are also added in Tab. S3.

S6. Drifting

As discussed in Sec. 5, we note some drifting when generating long videos interactively, both in our pipeline and in LongLive [7]. We hypothesize that this drifting is a result of no global sink and train-test environment mismatch as the model has originally been trained with a smaller attention window. We note that drifting, while present both in our pipeline and LongLive [7] in Fig. S5 can get worse in some samples with our pipeline (see Fig. S5 bottom) and can get better in other samples too (see Fig. S5 top). Faster inference with bigger attention windows would however reduce the necessitation of training with a smaller window size and remove any possible train-test distribution mismatch completely.

S7. Full Prompts

Prompts for all figures (top to bottom) are provided here.

For Fig. 1:

- A cheerful and playful Corgi running and frolicking in a sunlit park during the golden hour of sunset. The Corgi has a friendly smile, wagging tail, and bouncy gait as it runs through the grassy field. In the background, there are tall trees and families enjoying their evening out in the park. The scene begins with a close-up of the Corgi and gradually zooms out to reveal the expansive park and the setting sun casting a warm glow over everything. Medium to wide shot perspective.
- A stormtrooper from the Star Wars universe, clad in pristine white armor with a black helmet, is meticulously vacuuming a sandy beach. He bends down slightly, moving the vacuum cleaner back and forth across the sand with purposeful motions. His gloved hand firmly grips the handle of the vacuum as he navigates around rocks and debris. The sun sets behind him, casting long shadows and giving the scene a dramatic, golden glow. The background shows crashing waves and seagulls flying overhead. Medium close-up shot, focusing on the stormtrooper’s actions and the sweeping motion of the vacuum.

For Fig. 4:

- A movie trailer in a classic cinematic style, featuring the adventurous journey of a 30-year-old space man wearing a vibrant red wool knitted motorcycle helmet. The scene unfolds against a vast blue sky and a desolate salt desert landscape. Shot on 35mm film, the trailer showcases vivid and rich colors, capturing the hero as he navigates through the harsh terrain with determination. His helmet glints under

the sun, adding to the dramatic effect. The background is a mix of sweeping desert vistas and distant horizons, with the occasional shimmer of light reflecting off the salt flats. A dynamic medium shot with a sweeping overhead angle, emphasizing the hero’s resilience and the vastness of his adventure.

- A joyful, energetic golden retriever running freely across a lush green meadow. The dog has a playful expression, tongue hanging out, tail wagging enthusiastically as it runs. The sun shines brightly overhead, casting a warm glow over the scene. The grass sways gently in the breeze, adding to the serene and happy atmosphere. The camera follows the dog from a low angle, capturing its exuberant movement in a wide shot.
- An FPV drone shot capturing a majestic castle perched on a rocky cliff. The camera moves swiftly, revealing intricate stone walls, towering towers, and detailed gargoyles. The castle is partially shrouded in mist, adding a sense of mystery and grandeur. The cliff backdrop features jagged rocks and lush greenery, with patches of sunlight breaking through the clouds. The overall scene has a vivid and dynamic feel, with the camera angle emphasizing the height and imposing presence of the castle.
- Interactive Prompts:
 - Interior night, a dim dining room in a Chinese family home. A six-year-old Chinese girl with straight black bangs in a neat bob, warm almond eyes, a soft pink party dress with a small white bow at the waist, white ankle socks, red Mary Jane shoes, and a small paper birthday hat with tiny gold stars sits at a wooden table. In front of her is a frosted round birthday cake topped with five lit slim candles; the candlelight casts warm, flickering highlights on her cheeks and softly illuminates nearby plates, forks, and a checked tablecloth. Her Chinese mother—early 30s, warm almond eyes, neat black ponytail, simple silver stud earrings, cream cardigan over a light blue blouse—sits on her left, and her Chinese father—early 30s, short side-parted black hair, clean-shaven, dark-rim rectangular glasses, navy button-down with sleeves lightly rolled—sits on her right; both lean in close. Background falls to darkness with gentle bokeh from distant fairy lights. Wide establishing three-shot, locked tripod, 16:9 1920x1080, 24 fps, duration 4 s; lens 35 mm at f4 for modest depth of field; exposure biased to candle flame, warm key from candles at 2800 K with faint cool ambient 4800 K; natural motion blur; clean, realistic textures; no text, no watermark.
 - Mom and Dad grow quiet as the child closes her eyes and brings both hands together in front of her chest to make a silent wish; candle flames flutter in a faint air current and brighten momentarily. The girl is the same six-year-old Chinese child with straight black bangs, pink party dress with a white bow, white socks, red Mary Janes, and the starry paper hat. Her Chinese mother remains on her left, early 30s with a black ponytail, cream cardigan over a light blue blouse and silver studs; her Chinese father is on her right, early 30s with short side-parted black hair, clean-shaven, dark-rim glasses, and a navy button-down. Medium three-shot from across the table, slow 10 cm dolly-in, 16:9 1920x1080, 24 fps, duration 5 s; lens 50 mm at f2.8 for soft background; exposure locked to preserve candle detail; subtle handheld micro-shake; warm candle key with slight cool edge from a distant window; no text, no watermark.
 - Mom and Dad sing Happy Birthday together with the girl, smiling warmly as they lean in toward her; the girl looks between her parents and the cake. The girl keeps the same straight black bangs, pink party dress with white bow, white socks, red Mary Janes, and starry paper hat. Mom remains a Chinese woman in her early 30s with a neat black ponytail, cream cardigan over a light blue blouse, silver stud earrings; Dad is a Chinese man in his early 30s with short side-parted black hair, clean-shaven, dark-rim rectangular glasses, and a navy button-down. Close three-shot with emphasis on their faces and expressions; gentle rack focus among their faces, 16:9 1920x1080, 24 fps, duration 4 s; lens 85 mm at f2.0 for shallow depth of field; candlelight flicker visible on skin tones; noise-free low-light look; no text, no watermark.
 - The girl inhales and blows toward the candles in several steady breaths together with Mom and Dad, all three leaning in as the flames bend, stutter, and go out one by one until all five are extinguished. The frame extinguished. The room briefly darkens; a subtle nearby table lamp (3000 K) rises to a soft fill so faces remain visible. The girl is the same six-year-old Chinese child with straight black bangs, pink party dress with a white bow, white socks, red Mary Janes, and the starry paper hat. Her mother (early 30s, black ponytail, cream cardigan over light blue blouse, silver studs) and father (early 30s, short side-parted black hair, clean-shaven, dark-rim glasses, navy button-down) lean in close, blowing together. Close shot angled three-quarters with a slight push-in during the joint blow, 16:9 1920x1080, 24 fps, duration 4 s; lens 35 mm at f2.8; realistic motion blur on the exhales; no text, no watermark.
 - With the candles now out, Mom and Dad clap cheerfully, exchanging proud smiles, while the girl grins at them, eyes sparkling in the low, warm light. The child keeps the same straight black bangs, pink party dress with white bow, white socks, red Mary Janes, and starry paper hat. Mom remains early 30s, black ponytail, cream cardigan over a light blue blouse with silver stud earrings; Dad remains early 30s, short side-parted black hair, clean-shaven, dark-rim rectangular glasses, navy button-down. Medium close-up at the girl’s eye level, static, 16:9 1920x1080, 24 fps, duration 2 s; lens 50 mm at f2.2; consistent warm key from the lamp with minimal ambient; fine facial detail preserved; no

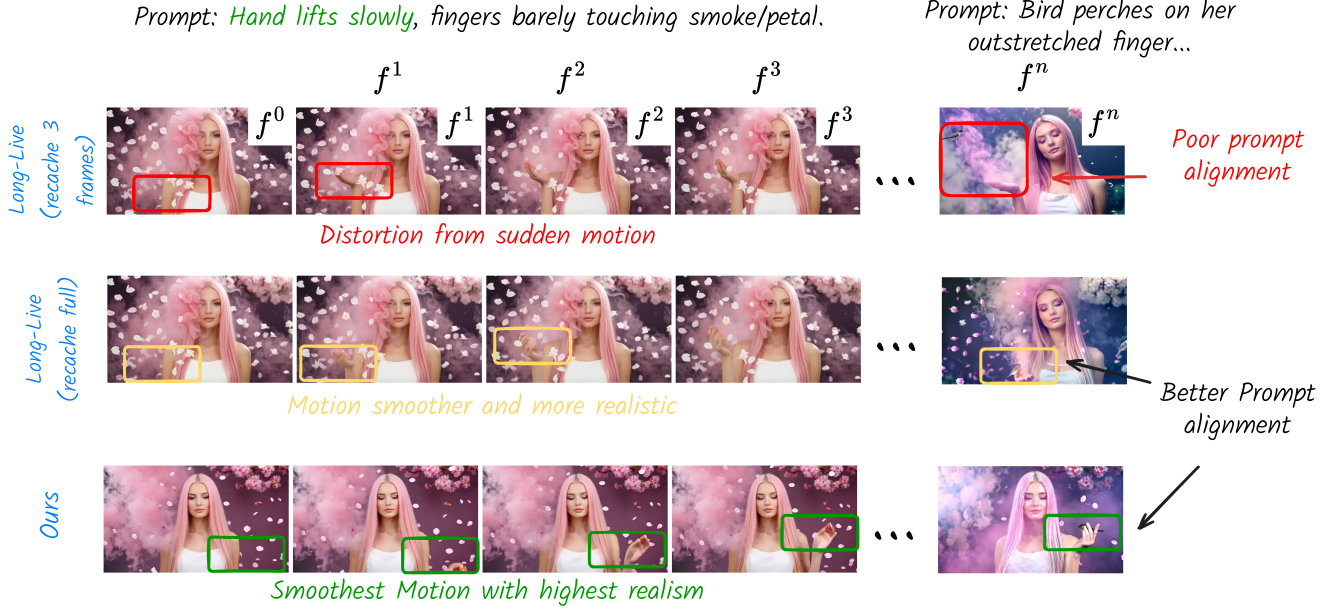


Figure S4. **KV recaching:** Recaching with fewer frames can lead to distortion from sudden motion and poor prompt alignment in generated videos.

Method	Background Consistency	Subject Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency	Quality Score	Semantic Score	Total Score
Wan2.1 [6]	0.98	0.95	0.99	0.98	0.65	0.67	0.66	0.97	0.86	0.95	0.90	0.74	0.51	0.22	0.25	0.27	0.85	0.80	0.84
FastVideo [9]	0.96	0.97	0.99	0.98	0.90	0.65	0.69	0.93	0.86	0.95	0.85	0.69	0.51	0.19	0.00	0.26	0.87	0.70	0.83
Causvid [8]	0.95	0.96	0.98	0.99	0.78	0.67	0.70	0.96	0.94	0.99	0.85	0.89	0.57	0.20	0.25	0.27	0.86	0.83	0.85
Rolling Forcing [5]	0.96	0.97	0.99	0.99	0.43	0.66	0.71	0.94	0.88	0.96	0.87	0.73	0.58	0.21	0.24	0.27	0.84	0.80	0.83
Self-Forcing [3]	0.96	0.96	0.99	0.98	0.69	0.66	0.69	0.94	0.87	0.97	0.89	0.83	0.62	0.21	0.24	0.27	0.85	0.82	0.84
Ours (w Self-Forcing ckpt)	0.96	0.95	0.99	0.99	0.61	0.65	0.69	0.93	0.89	0.97	0.85	0.76	0.57	0.20	0.24	0.27	0.84	0.80	0.84
LongLive [7]	0.97	0.97	0.99	0.99	0.43	0.67	0.69	0.96	0.88	0.97	0.90	0.81	0.59	0.21	0.24	0.27	0.84	0.82	0.83
Ours (w LongLive ckpt)	0.95	0.96	0.99	0.98	0.38	0.67	0.68	0.96	0.95	0.99	0.85	0.80	0.59	0.20	0.25	0.27	0.82	0.82	0.82
Krea-14B [1]	0.95	0.96	0.97	0.98	0.78	0.67	0.70	0.97	0.95	1.00	0.85	0.88	0.57	0.20	0.25	0.27	0.85	0.83	0.84
Ours (w Krea-14B ckpt)	0.95	0.96	0.98	0.98	0.78	0.67	0.71	0.97	0.95	1.00	0.84	0.89	0.59	0.20	0.25	0.27	0.85	0.83	0.85

Table S2. **VBench Scores:** VBench scores across all categories for competitor methods. Summary of scores in Tab. 1

Table S3. Quantitative analysis of ablative configurations

Configuration	Window (blocks)	Quality Score (\uparrow)	Semantic Score (\uparrow)	Total Score (\uparrow)
\mathcal{P}_1 (Self-Forcing)	7	0.8498	0.8206	0.8440
\mathcal{P}_2 (2-way, causal)	7	0.8454	0.8135	0.8390
\mathcal{P}_3 (3-way, causal)	7	0.8451	0.8149	0.8390
\mathcal{P}_4 (5-way, causal)	7	0.8458	0.8109	0.8388
\mathcal{P}_3 (bidirectional)	3	0.8388	0.8056	0.8321
\mathcal{P}_3 (bidirectional)	4	0.8437	0.8155	0.8381
\mathcal{P}_4 (bidirectional)	5	0.8460	0.8107	0.8389
\mathcal{P}_4 (bidirectional)	6	0.8455	0.8112	0.8386
Ours (\mathcal{P}_4 , bidirectional)	7	0.8435	0.8024	0.8353

text, no watermark.

- Mom gently pats the girl’s head, smoothing a strand of her straight black bangs, while Dad rests and lightly taps his right hand on the girl’s near shoulder in a reassuring gesture. The girl—six-year-old Chinese child in a soft pink party dress with a white bow, white socks, red Mary Janes, and a starry paper hat—giggles

Table S4. Latency and Drifting

Method	Quality Score	Semantic Score	Total Score	First Frame Latency (s)	Quality Drift
Self-Forcing	0.8295	0.7773	0.8191	0.52	6.54
Rolling-Forcing	0.8349	0.7960	0.8271	1.59	2.77
LongLive	0.8328	0.8056	0.8274	0.55	2.87
Ours (LongLive)	0.8314	0.8163	0.8283	0.74	2.80

softly. Mom keeps her neat black ponytail, cream cardigan over a light blue blouse, silver studs; Dad keeps his short side-parted black hair, clean-shaven look, dark-rim rectangular glasses, and navy button-down. Close three-shot emphasizing hands and expressions; subtle rack focus from Mom’s pat to Dad’s shoulder touch to the girl’s smiling face, 16:9 1920x1080, 24 fps, duration 4 s; lens 85 mm at f2.0 for shallow depth of field; warm, soft fill from the table lamp; no text, no watermark.

For Fig. S4:

- A beautiful model with long, flowing pink hair partially covered by swirling pink smoke. She has delicate features and a serene expression, standing gracefully



Figure S5. **Drifting:** Long interactive video generation without a sink can cause drifting in generated videos.

against a backdrop of gently falling sakura petals. The petals float softly in the air, creating a dreamy and ethereal atmosphere. The model is dressed in a simple, elegant white gown that complements the soft pink hues of the smoke. The scene is captured in a medium close-up shot, emphasizing the intricate details of the pink smoke and the gentle beauty of the sakura petals.

- A beautiful model with long, flowing pink hair partially covered by swirling pink smoke. She has delicate features and a serene expression, standing gracefully against a backdrop of gently falling sakura petals. The petals float softly in the air, creating a dreamy and ethereal atmosphere. The model lifts her hand slowly, as if caressing a petal, her fingers barely touching the delicate pink smoke. The scene is captured in a medium close-up shot, emphasizing the intricate details of the pink smoke and the gentle beauty of the sakura petals.
- A beautiful model with long, flowing pink hair partially covered by swirling pink smoke. She has delicate features and a serene expression, standing gracefully against a backdrop of gently falling sakura petals. The petals float softly in the air, creating a dreamy and ethereal atmosphere. The model lifts her hand slowly, as if caressing a petal, her fingers barely touching the delicate pink smoke, while a soft breeze causes a few more petals to flutter around her. The scene is captured in a medium close-up shot, emphasizing the intricate details of the pink smoke and the gentle beauty of the sakura petals.
- A beautiful model with long, flowing pink hair partially covered by swirling pink smoke. She has delicate features and a serene expression, standing gracefully against a backdrop of gently falling sakura petals. The petals float softly in the air, creating a dreamy and ethereal atmosphere. The model lifts her hand slowly, as if caressing a petal, her fingers barely touching the delicate pink smoke, while a soft breeze causes a few more petals to flutter around her. She gently closes her eyes, her lashes resting softly as the serene scene remains undisturbed. The scene is captured in a medium close-up shot, emphasizing the intricate details of the pink smoke and the gentle beauty of the sakura petals.
- A beautiful model with long, flowing pink hair partially covered by swirling pink smoke. She has delicate features and a serene expression, standing gracefully against a backdrop of gently falling sakura petals. The petals float softly in the air, creating a dreamy and ethereal atmosphere. The model lifts her hand slowly, as if caressing a petal, her fingers barely touching the delicate pink smoke, while a soft breeze causes a few more petals to flutter around her. Suddenly, she takes a step forward, her movement barely disturbing the serene scene, and a small bird flits to a nearby branch and settles there. Medium close-up.
- A beautiful model with long, flowing pink hair partially covered by swirling pink smoke. She has delicate features and a serene expression, standing gracefully against a backdrop of gently falling sakura petals. The petals float softly in the air, creating a dreamy and ethereal atmosphere. The model lifts her hand slowly, as if caressing a petal, her fingers barely touching the delicate pink smoke, while a soft breeze causes a few more petals to flutter around her. Suddenly, she takes a step forward, her movement barely disturbing the serene scene, and a small bird perches delicately on her outstretched finger. The pink and cyan-blue smoke gradually grows denser, softly enveloping the entire scene. Medium close-up.

For Fig. S6:

- A close-up shot of a fluffy gray cat with green eyes eating food from a white ceramic bowl. The cat has a curious expression and its tail is gently swishing behind it. The bowl is placed on a wooden table, and the cat's whiskers are brushing against the surface of the bowl as it eats. The lighting is soft and warm, casting gentle shadows on the table. The cat's natural movements include licking its lips and occasionally pausing to look around. Medium shot focusing on the interaction between the cat and the bowl.
- A breathtaking fantasy landscape featuring towering ancient trees with glowing leaves, surrounded by mystical floating islands and cascading waterfalls. Enchanted forests filled with luminescent flora and fauna, and distant, majestic mountains under a starlit sky. The ground is covered in soft, shimmering grass that glows gently. The scene is bathed in an ethereal, magical light, creating an otherworldly atmosphere. Wide shot, static scene.
- A confused panda sitting in a classroom filled with desks and chairs, surrounded by other animated animal students taking notes. The panda is holding a pencil and staring at a complex calculus equation on a chalkboard, scratching its head with a puzzled expression. The classroom has typical school decor including posters

on the walls and a teacher's desk at the front. The panda looks lost and overwhelmed, trying to understand the mathematical concepts being taught. Medium shot focusing on the panda's reaction and the chalkboard in the background.

- A close-up view of a cluster of vibrant green grapes on a rotating glass table under soft, diffused lighting. The grapes are large and plump, reflecting the gentle light as they rotate slowly, showcasing their smooth, shiny surfaces. The background is blurred, focusing attention solely on the grapes and their subtle reflections. The camera remains static, capturing the serene and detailed motion of the rotating grapes. Close-up shot.
- A futuristic super robot standing tall and vigilant, protecting a bustling city skyline from a looming threat. The robot has a sleek, metallic design with glowing blue energy lines running across its body. It stands in a heroic pose, arms raised, ready to defend the city below. The urban landscape features towering skyscrapers, bustling streets, and neon lights reflecting off the wet pavement after a rain shower. The scene is captured in a sweeping wide shot, showcasing the robot's imposing size and the vibrant city life behind it.

For Fig. S7:

- Interactive Prompts:
 - A somber, gloomy atmosphere envelops the dimly lit, murky river beneath a crumbling overpass. Brownish silt churns through the water, diffusing the scant light leaking from distant streetlamps. In the foreground, a gaunt man in his mid-forties, his thin, greasy black hair plastered to his skull, struggles against the current; his sunken cheeks and bloodshot eyes betray years of drink. A tattered gray shirt and ripped trousers flap around his skeletal frame as he thrashes, desperate for air. Rusted cans, frayed plastic bags, and strands of river weed drift past, hinting at the surrounding urban decay. Shadows dance across his hollow features, deepening the suffocating mood. The shot never cuts, the lens trembling slightly with the current. medium close-up
 - A somber, gloomy atmosphere envelops the dimly lit, murky river beneath a crumbling overpass. Brownish silt churns through the water, diffusing the scant light leaking from distant streetlamps. In the foreground, the same gaunt man in his mid-forties, his thin, greasy black hair plastered to his skull, stretches a trembling hand upward, fingertips slicing the turbid flow as though searching for the surface. His tattered gray shirt and ripped trousers cling to his skeletal frame, accentuating the hollow of his collarbones. Rusted cans, frayed plastic bags, and tangled river weed drift past, testifying to the city's neglect. Shadows oscillate across his drawn face, intensifying the oppressive dread, while the unbroken shot quivers gently with the current. medium close-up
 - A somber, gloomy atmosphere envelops the dimly lit, murky river beneath a crumbling overpass. Brownish silt churns through the water, diffusing the scant light leaking from distant streetlamps. In the foreground, the same gaunt man in his mid-forties, his thin, greasy black hair plastered to his skull, suddenly clutches his chest as a violent spasm contorts his emaciated torso, forcing bubbles from his gaping mouth. His tattered gray shirt and ripped trousers constrict around protruding ribs. Rusted cans, frayed plastic bags, and tangled river weed drift past, echoing the city's decay. Shadows ripple across his mask-like face, heightening the stifling despair, while the uninterrupted shot shivers subtly with the restless current. medium close-up
 - A somber, gloomy atmosphere envelops the dimly lit, murky river beneath a crumbling overpass. Brownish silt churns through the water, diffusing the scant light leaking from distant streetlamps. In the foreground, the same gaunt man in his mid-forties, his thin, greasy black hair plastered to his skull, slowly turns his hollow eyes to follow a jagged shard of glass drifting by, its edge catching the faint glow like a ghostly beacon. His tattered gray shirt and ripped trousers ripple against his skeletal limbs. Rusted cans, frayed plastic bags, and tangled river weed orbit him, symbols of urban neglect. Shadows glide over his strained features, compounding the claustrophobic menace, while the continuous shot quakes minutely within the current. medium close-up
 - A somber, gloomy atmosphere envelops the dimly lit, murky river beneath a crumbling overpass. Brownish silt churns through the water, diffusing the scant light leaking from distant streetlamps. In the foreground, the same gaunt man in his mid-forties, his thin, greasy black hair plastered to his skull, fixes his hollow gaze straight ahead and throws his head back, mouth wide in ragged, desperate gasps that erupt in frantic chains of bubbles. His tattered gray shirt and ripped trousers cling to his angular frame. Rusted cans, frayed plastic bags, and tangled river weed drift past, painting a portrait of civic disregard. Shadows flicker across his strained face, thickening the air of doom, while the unbroken shot trembles faintly with the relentless current. medium close-up
 - A somber, gloomy atmosphere envelops the dimly lit, murky river beneath a crumbling overpass. Brownish silt churns through the water, diffusing the scant light leaking from distant streetlamps. In the foreground, the same gaunt man in his mid-forties, his thin, greasy black hair plastered to his skull, slowly closes his swollen eyelids and becomes eerily motionless, suspended like a ragged effigy in the gloom. His tattered gray shirt and ripped trousers float around his skeletal limbs, barely stirring. Rusted cans, frayed plastic bags, and tangled river weed drift past, silent witnesses to urban blight. Shadows settle upon his pallid face, sealing the suffocating tension, while the continuous shot vibrates softly within the oppressive current. medium close-up
- Interactive Prompts:

Prompt: A close-up shot of a fluffy gray cat with green eyes eating food from a white ceramic bowl. The cat has a curious expression and its tail is gently swishing behind it. The bowl is ...

Krea-14B



Ours



Prompt: A breathtaking fantasy landscape featuring towering ancient trees with glowing leaves, surrounded by mystical floating islands and cascading waterfalls. Enchanted forests filled ...

Krea-14B

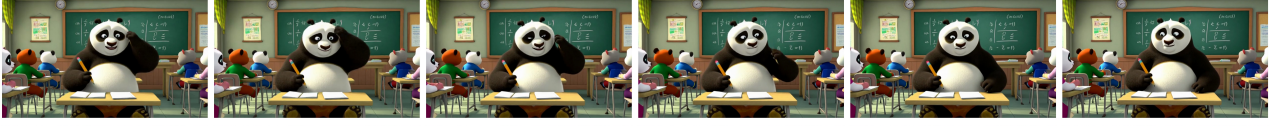


Ours



Prompt: A confused panda sitting in a classroom filled with desks and chairs, surrounded by other animated animal students taking notes. The panda is holding a pencil and staring at a ...

Krea-14B

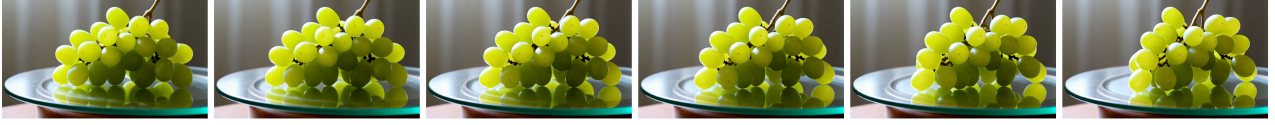


Ours

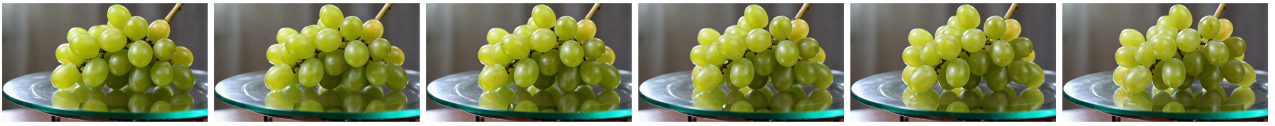


Prompt: A close-up view of a cluster of vibrant green grapes on a rotating glass table under soft, diffused lighting. The grapes are large and plump, reflecting the gentle light as they rotate ...

Krea-14B



Ours



Prompt: A futuristic super robot standing tall and vigilant, protecting a bustling city skyline from a looming threat. The robot has a sleek, metallic design with glowing blue energy lines ...

Krea-14B



Ours



Figure S6. **Qualitative Comparisons:** Comparing short video generation with Krea-14B[1]

– A realistic video of a Texas Hold'em poker event at a casino. A male player in his late 30s with a medium build, short dark hair, light stubble, and a sharp jawline wears a fitted navy blazer over a charcoal crew-neck tee, dark jeans, and a stainless-steel watch. He sits at a well-lit poker table and tightly grips his hole cards, wearing a tense, serious expression. The table is filled with chips of various colors, the dealer is seen dealing cards, and several rows of slot machines glow in the background. The camera focuses on the player's strained

concentration. Wide shot to medium close-up.

– A realistic video of a Texas Hold'em poker event at a casino. The same male player—late 30s, medium build, short dark hair, light stubble, sharp jawline—dressed in a fitted navy blazer over a charcoal tee, dark jeans, and a stainless-steel watch—flicks his cards onto the felt, then leans back in the chair with arms spread wide in celebration. The dealer continues dealing to the table as stacks of multicolored chips crowd the surface; slot machines and nearby



Figure S7. **Qualitative Comparisons:** Comparing long interactive video generation with LongLive[7]

- patrons fill the background. The camera locks onto the player's exuberant reaction. Wide shot to medium close-up.
- A realistic video of a Texas Hold'em poker event at a casino. The same late-30s male player, medium build with short dark hair and light stubble, wearing a navy blazer, charcoal tee, dark jeans, and a stainless-steel watch, reveals the winning hand and leans back in celebration while the dealer keeps the game moving. A nearby patron claps and cheers for the winner, amplifying the festive atmosphere. The table brims with colorful chips, with slot machines and other tables behind. The camera centers on the winner's reaction as the applause rises. Wide shot to medium close-up.
- A realistic video of a Texas Hold'em poker event at a casino. The same male player—late 30s, medium build, short dark hair, light stubble—still in his navy blazer, charcoal tee, dark jeans, and stainless-steel watch—sits upright and begins neatly arranging the stacks of chips in front of him, methodically straightening and organizing the piles. The dealer continues dealing, and rows of slot machines pulse in the background. The camera captures the composed, pur-

- poseful movements at the well-lit table. Wide shot to medium close-up.
- A realistic video of a Texas Hold’em poker event at a casino. The same late-30s male player with short dark hair, light stubble, and a sharp jawline, wearing a fitted navy blazer over a charcoal tee, dark jeans, and a stainless-steel watch, glances over his chips and breaks into a proud, self-assured smile, basking in the victorious moment. Multicolored chips crowd the felt, the dealer works the table, and slot machines glow behind. The camera emphasizes the winner’s pride and satisfaction. Wide shot to medium close-up.
 - A realistic video of a Texas Hold’em poker event at a casino. The same male player—late 30s, medium build, short dark hair, light stubble—dressed in a navy blazer, charcoal tee, dark jeans, and a stainless-steel watch—shares a celebratory high-five with a nearby patron after the win, laughter and cheers rippling around the table. Stacks of chips are spread across the felt, the dealer continues dealing, and the background features rows of slot machines and other patrons. The camera focuses on the jubilant interaction. Wide shot to medium close-up.
- Interactive Prompts:
 - A vibrant Christmas celebration in Rio de Janeiro, Brazil. The scene features a lively street filled with colorful lights, festive decorations, and cheerful locals in casual summer attire. At the center of the street stands a slightly overweight Black adult man in a blue shirt, holding a small red ball in his hand. People are walking, dancing, and enjoying the festive atmosphere. In the background, there are tall buildings, palm trees, and a glimpse of the iconic Christ the Redeemer statue. The sun is setting, casting a warm golden glow over the scene. Medium shot capturing the bustling energy of the street.
 - A vibrant Christmas celebration in Rio de Janeiro, Brazil. The scene features a lively street filled with colorful lights, festive decorations, and cheerful locals in casual summer attire. A slightly overweight Black adult man in a blue shirt tosses a small red ball into the air, drawing a small crowd of children and adults who watch in awe. The sun is setting, casting a warm golden glow over the scene. Medium shot capturing the bustling energy of the street.
 - A vibrant Christmas celebration in Rio de Janeiro, Brazil. The scene features a lively street filled with colorful lights, festive decorations, and cheerful locals in casual summer attire. The slightly overweight Black adult man in a blue shirt switches to flowing ribbons, twirling long, colorful streamers that ripple through the warm air, eliciting gasps and cheers from the growing crowd of children and adults. The sun is setting, casting a warm golden glow over the scene. Medium shot capturing the bustling energy of the street.
 - A vibrant Christmas celebration in Rio de Janeiro, Brazil. The scene features a lively street filled with colorful lights, festive decorations, and cheerful locals in casual summer attire. The slightly overweight Black adult man in a blue shirt hold a flaming torch and continues performing with it, sending the crowd’s excitement surging higher as they watch in awe. The sun is setting, casting a warm golden glow over the scene. Medium shot capturing the bustling energy of the street.
 - A vibrant Christmas celebration in Rio de Janeiro, Brazil. The scene features a lively street filled with colorful lights, festive decorations, and cheerful locals in casual summer attire. The slightly overweight Black adult man in a blue shirt balances the flaming torch on his chin, while children press closer to the front, eyes wide with wonder. The sun is setting, casting a warm golden glow over the scene. Medium shot capturing the bustling energy of the street.
 - A vibrant Christmas celebration in Rio de Janeiro, Brazil. The scene features a lively street filled with colorful lights, festive decorations, and cheerful locals in casual summer attire. The slightly overweight Black adult man in a blue shirt twirls the flaming torch between his fingers, and the audience begins applauding, waves of clapping rolling through the crowd. The sun is setting, casting a warm golden glow over the scene. Medium shot capturing the bustling energy of the street.
 - Interactive Prompts:
 - A woman walks slowly alongside her boyfriend on a picturesque suburban street at sunset. The woman—mid-20s, shoulder-length dark hair, soft features, light natural makeup—wears a pastel light spring coat over a white blouse, slim jeans, and simple white sneakers. The man—late-20s, short neatly styled dark hair with faint stubble and a warm smile—wears a sky-blue oxford shirt under a heather-gray sweater vest, tan chinos, and clean lace-up sneakers, with a slim leather watch. Both have relaxed expressions, holding hands as they stroll together. The background shows a row of houses with blooming flowers and lush green lawns. The sun casts a warm golden glow, creating long shadows behind them. Medium shot, focusing on their interaction and the serene environment around them.
 - A woman walks slowly alongside her boyfriend on a picturesque suburban street at sunset. The woman—mid-20s, pastel light spring coat, white blouse, slim jeans, white sneakers, stops to pick a white flower from a garden border; her coat sways slightly as she gently brings it to her nose to inhale its fragrance. The 20s-man—sky-blue oxford under a heather-gray sweater vest, tan chinos, clean sneakers, leather watch—waits patiently, smiling as she admires the bloom. The background remains a row of houses with blooming flowers and lush green lawns. The sun’s warm golden glow continues to cast long shadows behind them. Medium shot, focusing on their interaction and the serene environment.
 - A woman walks slowly alongside her boyfriend on a picturesque suburban street at sunset. After smelling the white flower, the same woman—pastel coat, white blouse, slim jeans, white sneakers—first lifts the bloom up playfully, then places it behind her ear, tucking it securely in place. The same man—sky-blue oxford shirt beneath a heather-gray sweater vest, tan chinos, clean sneakers, leather watch—watches with a warm smile. The background remains a row of houses with blooming flowers and lush green lawns. The sun’s warm golden glow continues to cast long shadows behind them. Medium shot, focusing on their interaction and the serene environment.
 - A woman walks slowly alongside her boyfriend on a picturesque suburban street at sunset. The woman—mid-20s, pastel light spring coat, white blouse, slim jeans, white sneakers—now with the white flower tucked behind her ear, smiles as a playful breeze ruffles a few strands of her shoulder-length dark hair and rustles the leaves of a nearby tree. The same man—sky-blue oxford under a heather-gray sweater vest, tan chinos, clean sneakers, leather watch—looks on with an affectionate expression. The background remains a row of houses with blooming flowers and lush green lawns. The sun’s warm golden glow continues to cast long shadows behind them. Medium shot, focusing on their interaction and the serene atmosphere.
 - A woman walks slowly alongside her boyfriend on a picturesque suburban street at sunset. The white flower remains tucked behind the mid-20s-woman’s ear; her pastel coat catch the light as they pause, gazing into each other’s eyes with warm smiles and begin speaking softly, exchanging quiet words. The same man—sky-blue oxford beneath a heather-gray sweater vest, tan chinos, clean sneakers, leather watch—stands close, his neatly styled dark hair outlined by the golden rim light. A playful breeze rustles the leaves of a nearby tree, adding a gentle sound to the serene atmosphere. The background remains a row of houses with blooming flowers and lush green lawns. The sun’s warm golden glow continues to cast long shadows behind them. Medium shot, focusing on their tender exchange and the peaceful surroundings.
 - A woman walks slowly alongside her boyfriend on a picturesque suburban street at sunset. The white flower stays tucked behind the mid-20s-woman’s ear as they intertwine their fingers once again and resume their stroll together; her pastel light spring coat drapes softly over her white blouse and slim jeans, white sneakers moving in calm rhythm. The same man—sky-blue oxford under a heather-gray sweater vest, tan chinos, clean sneakers, leather watch—walks in step beside her, faint stubble catching the light. A playful breeze rustles the leaves of a nearby tree, adding a gentle sound to the serene atmosphere. The background remains a row of houses with blooming flowers and lush green lawns. The sun’s warm golden glow continues to cast long shadows behind them. Medium shot, focusing on their renewed connection and the tranquil environment.

References

- [1] Krea AI. Krea realtime 14b: Real-time video generation. 2025. 3, 5
- [2] Anthropic. Claude 4.5 sonnet. <https://www.anthropic.com>, 2024. 1
- [3] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. 1, 3
- [4] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 2
- [5] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025. 3
- [6] 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. 3
- [7] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong

Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025. [2](#), [3](#), [6](#)

- [8] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025. [3](#)
- [9] Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Vsa: Faster video diffusion with trainable sparse attention. *arXiv preprint arXiv:2505.13389*, 2025. [3](#)