Generating, Fast and Slow: Scalable Parallel Video Generation with Video Interface Networks

Supplementary Material

All videos shown in the manuscript can be found in the *index.html* file of the supplementary material.

6. Glossary of Mathematical Notations

We summarize notations used in the paper in Table 3.

Notations	Description
T_{emb}	Text Embeddings
t_{emb}	Time Embeddings
N	Total Tokens in Input Video
$X^{1:N}$	Input Video with N tokens
$X_t^{1:N,T_s}$	Input Video sub-sampled every T_s seconds
$Z_{init}^{1:N_{global}}$	Initial N_{Global} Global Tokens
Z_t	Final Global Tokens at step t
N_s	Tokens in F frames of the video
$X_t^{iN_s:(i+1)N_s}$	i^{th} Temporal Chunk of Input Video
N_{local}	Tokens in the last F_{local} frames of a chunk
$X_t^{i,N_{local}}$	Last F_{local} frames of the i^{th} chunk

Table 3. Glossary of mathematical notations utilized in Section 3.

7. FLOPs Breakdown

We further stratified the FLOPs analysis in Fig. 10 over different operations in the architecture. We present our results in Fig. 13 comparing full attention on the DiT to the VIN-DiT ensemble. VIN and Full-attention have similar QKV and feed-forward costs due to nearly comparable token counts in the input video. The VIN module incurs marginal overhead from processing the global tokens, but also enables a significant reduction in the total attention cost by replacing $O(N^2)$ attention on a long N-token video with multiple local attentions over shorter chunks.

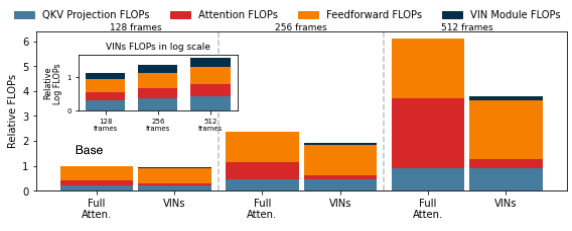

Figure 13. FLOPs breakdown comparison of VINs against Full Attention

8. Comparison to Open Weight Models

We compared our approach to three open weight DIT models viz. OpenSora v1.2 [55], Mochi-1 [45] and Hunyuan-Video [11] using full attention on the long video. We analyzed videos generated at the 256 frame setting on different VBench metrics in Table 2. While OpenSora has been natively trained to generate longer videos, both Mochi-1 and HunyuanVideo were evaluated at the extended frame setting where inference was performed beyond the recommended number of frames. As a strong baseline, we also report the results from using our base model in the extended setting. Our primary finding was that at longer frame settings, all the DiT models exhibit stagnated motion in the video where temporal dynamics often become near static. As a result, while this leads to high consistency scores, the dynamic degree suffers significantly. This corroborates our findings in the main section where our base model also exhibits this detrimental tradeoff. Moreover, VIN not only maintains consistency at par with the open models but also outperforms them on the dynamic degree metric by a significant margin. We also noticed that outputs from both Mochi-1 and Hunyuan video qualitatively deteriorate at longer frame settings, with the latter also showing temporal flickering artifacts. Fig. 15 and 16 show the qualitative comparisons between models.

9. Additional VBench Evaluation Metrics

We also evaluated generated videos from different chunk based methods on two other VBench [21] metrics, {Text-Video Alignment, Motion Smoothness, Dynamic Degree}, as shown in Fig. 14. We observed the following:

1. **Text-Video Alignment:** Alignment scores, as measured by ViCLIP [50], were primarily bifurcated by the base model. StreamingT2V [18] that uses a ModelScope [48] base has a considerably lower score than the other methods, which use a common base model. Overall, text-video alignment scores diminish with video length, with the autoregression scores deteriorating more than the other parallel inference and full-generation methods.
2. **Motion Smoothness:** VBench measures motion smoothness via motion priors obtained from a frame interpolation model [28]. We observe that the motion smoothness of our model is uniformly high across all evaluations. In principle, this metric captures the essence of pixel-level motion fluidity similar to our experiment in Section 4.4; however, it fundamentally differs in the motion-prior we utilize. The AMT model, on which the

Model	Extended Setting	Subj. Consist.	BG Consist.	Imag. Qual.	Temp. Flicker.	Motion Smooth.	Dyna. Degree	Text-Video Align.
OpenSora v1.2 [55]	No	96.75%	97.61%	56.85%	99.53%	98.50%	63.34%	**26.85%**
Mochi-1 [45]	Yes (163 → 256)	96.99%	97.28%	56.94%	99.40%	97.02%	60.64%	25.15%
HunyuanVideo [11]	Yes (128 → 256)	**97.37** %	**97.76** %	54.39 %	98.32%	97.96%	70.83%	26.44%
Base DiT (Ours)	Yes (128 → 256)	96.91 %	97.20 %	**63.52** %	**99.64%**	**98.57%**	68.15%	25.57%
VINs (Ours)	No	96.4%	97.13%	58.5 %	99.12%	**98.55** %	**87.32** %	25.43%

Table 2. Comparison of VINs against Open Weight DiT Models. We considered three popular open weight models and evaluated them on key VBench metrics at the 256-frame settings. Inference was performed via the Full attention mode. While most models exhibit good consistency at long frames, they tend to often stagnate and become static, leading to poor dynamic degree scores.

VBench metric is based, uses bidirectional correlation volumes and regresses on frame interpolation and, therefore, reasons about coarser flows in the video. On the other hand, we use the RAFT [46] model that is trained to measure the optical flow of each pixel and gives a more faithful measure of finer flows across frames. The difference in motion smoothness measured by the former is, therefore, less apparent compared to RAFT, where the difference is much more pronounced.

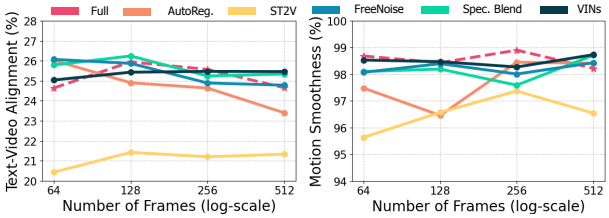

Figure 14. VBench evaluation results on Text-Video alignment, motion smoothness, and dynamic degree of the video.

10. Qualitative Visualization

Figs. 17-21 demonstrate videos generated by VINs and other methods considered in this paper over different prompts. Figs. 17, 18, 19 show the generations on prompts that primarily consist of diverse objects or background details where semantic coherence is crucial. We observe that VINs maintain the subject, object, and background consistency while also generating more photorealistic videos. We also notice the subject-aware softening mentioned in Section 4.3. This was specifically observed where the prompt's background description was missing. For example, "A Raccoon Dressed in a Suit Playing the Trumpet," and "Grizzly Bear Trying to Learn Calculus." Figs. 20 and 21 show the generated videos on landscape prompts along with their y-t cross-sections. As evidenced by the latter visualizations, the perturbations across chunk boundaries for VINs are minimal and the transitions are smoother. As highlighted in Fig. 10 (a), one can also qualitatively observe the disparity of the consistency-dynamic degree tradeoff between full attention and VINs. Full attention methods tend to generate outputs with undesirable artifacts more often than not. These frequently manifest as reduced motion, diminished object complexity, and loopy distortions, exemplified by the "Grizzly bear," "Dog wearing a cape," "Rabbit in a fantasy landscape," and "Shark in the ocean" prompts.

Figure 15. Qualitative comparison of VINs against open weight models at 256 frames.

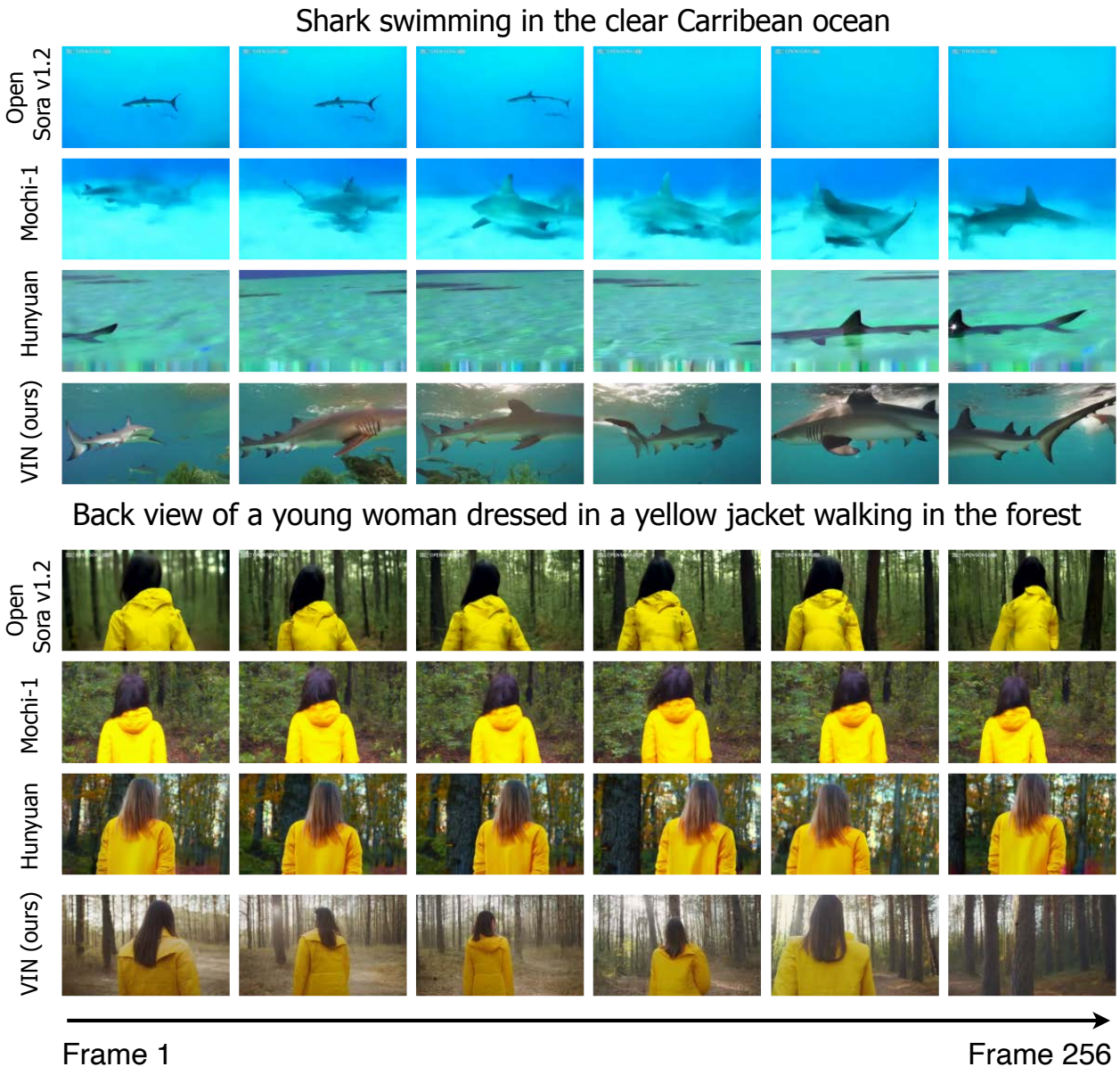

Figure 16. Qualitative comparison of VINs against open weight models at 256 frames.

Figure 17. Qualitative visualizations across different methods.

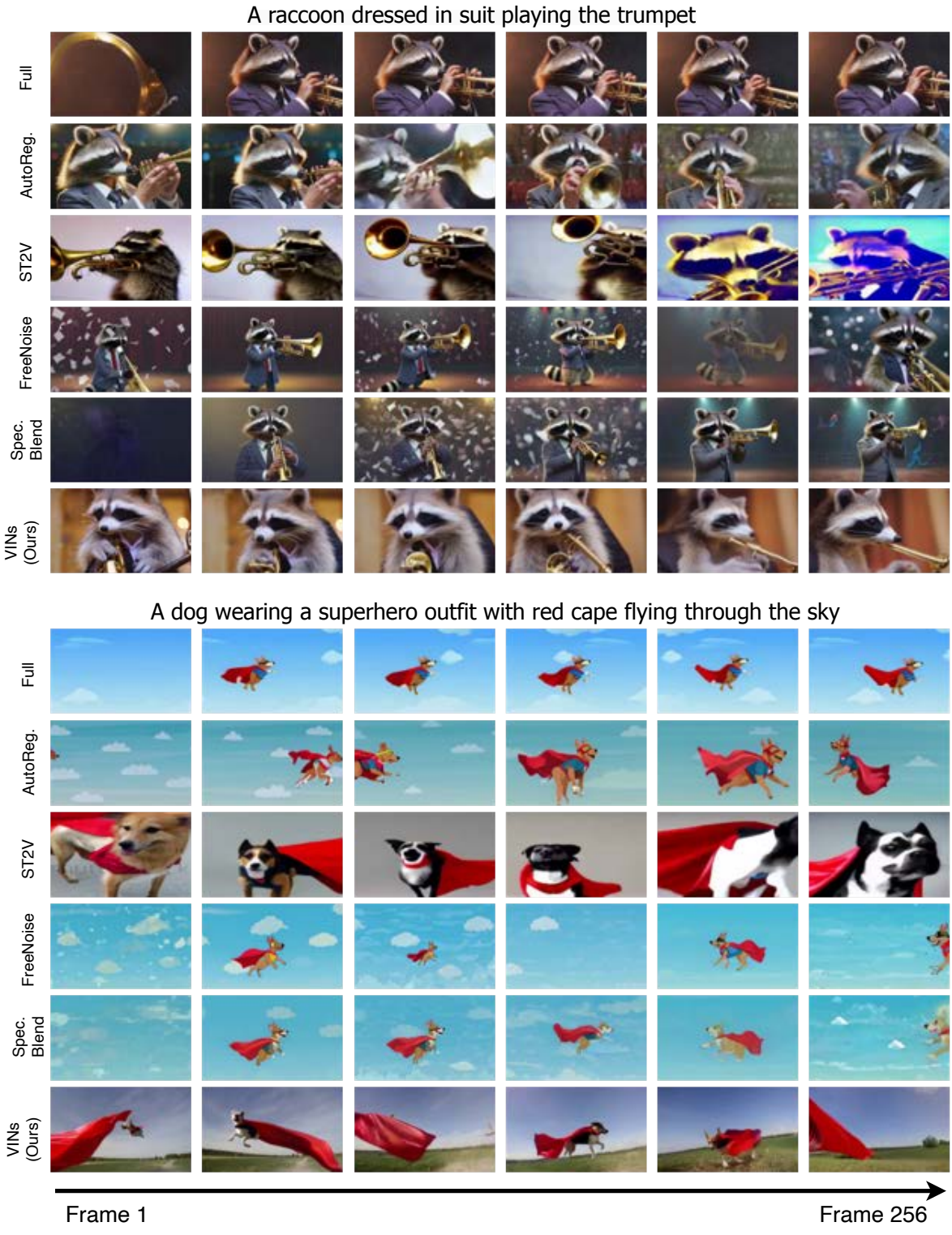

Figure 18. Qualitative visualizations across different methods.

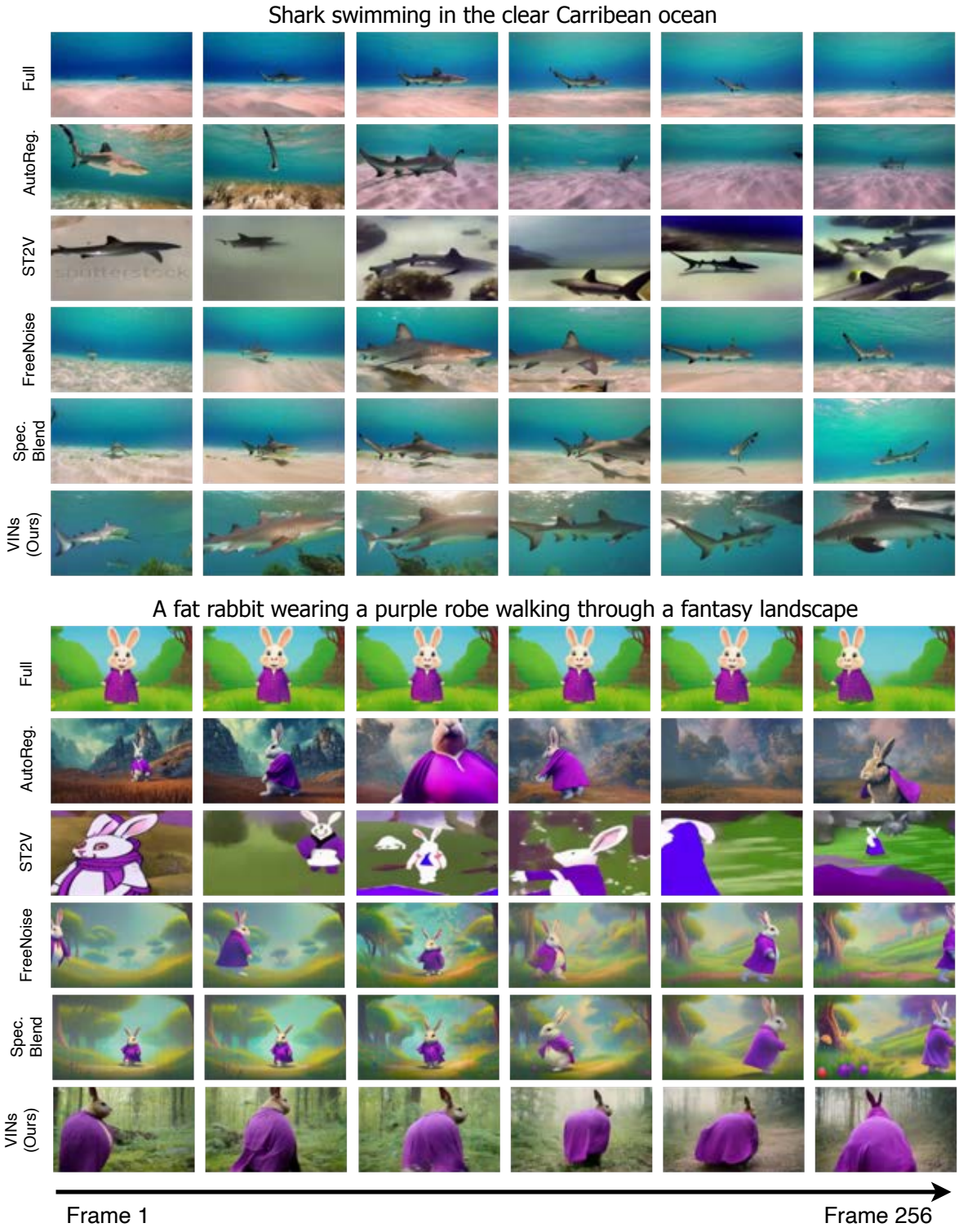

Figure 19. Qualitative visualizations across different methods.

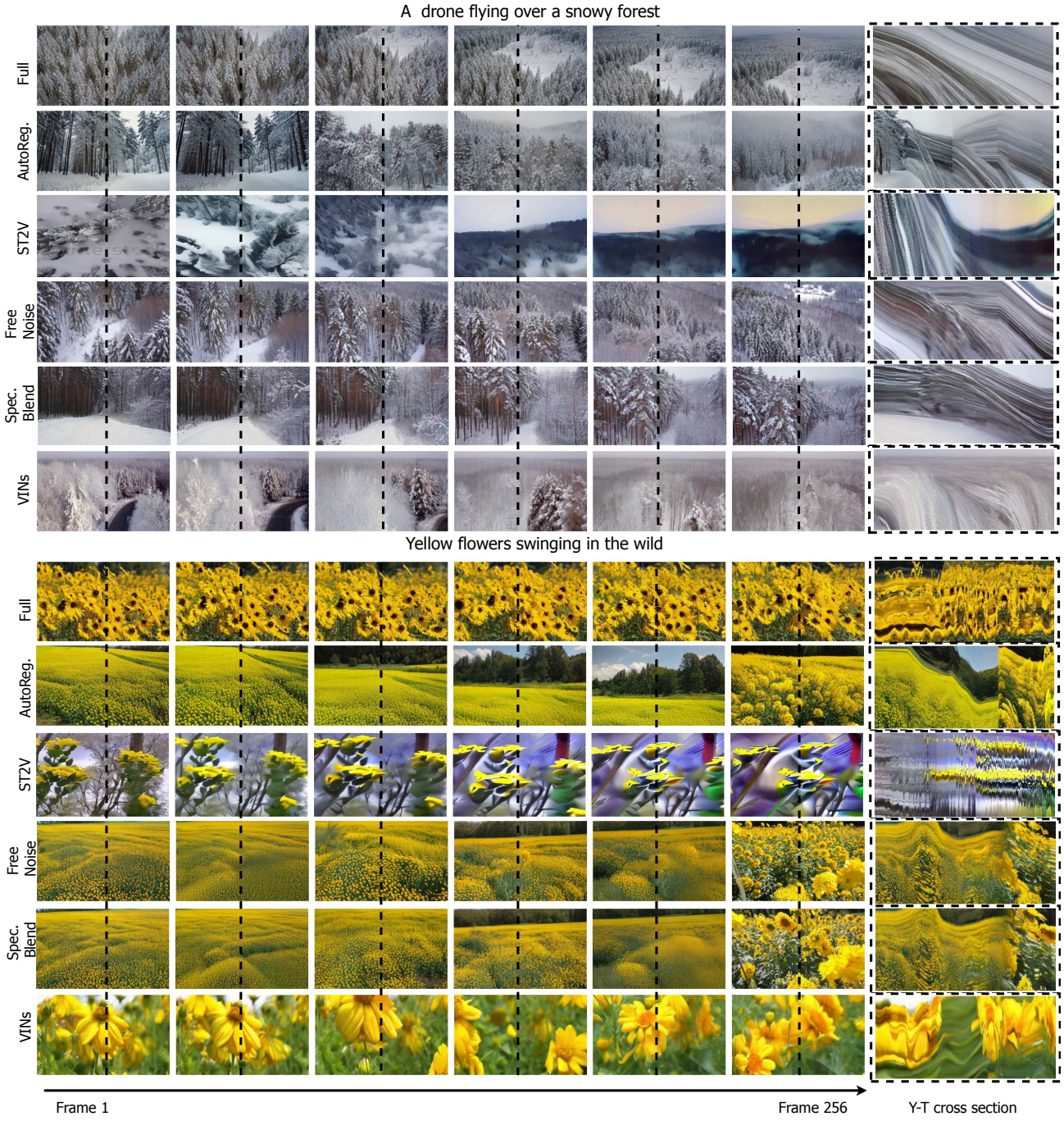

Figure 20. Qualitative visualizations across different methods.

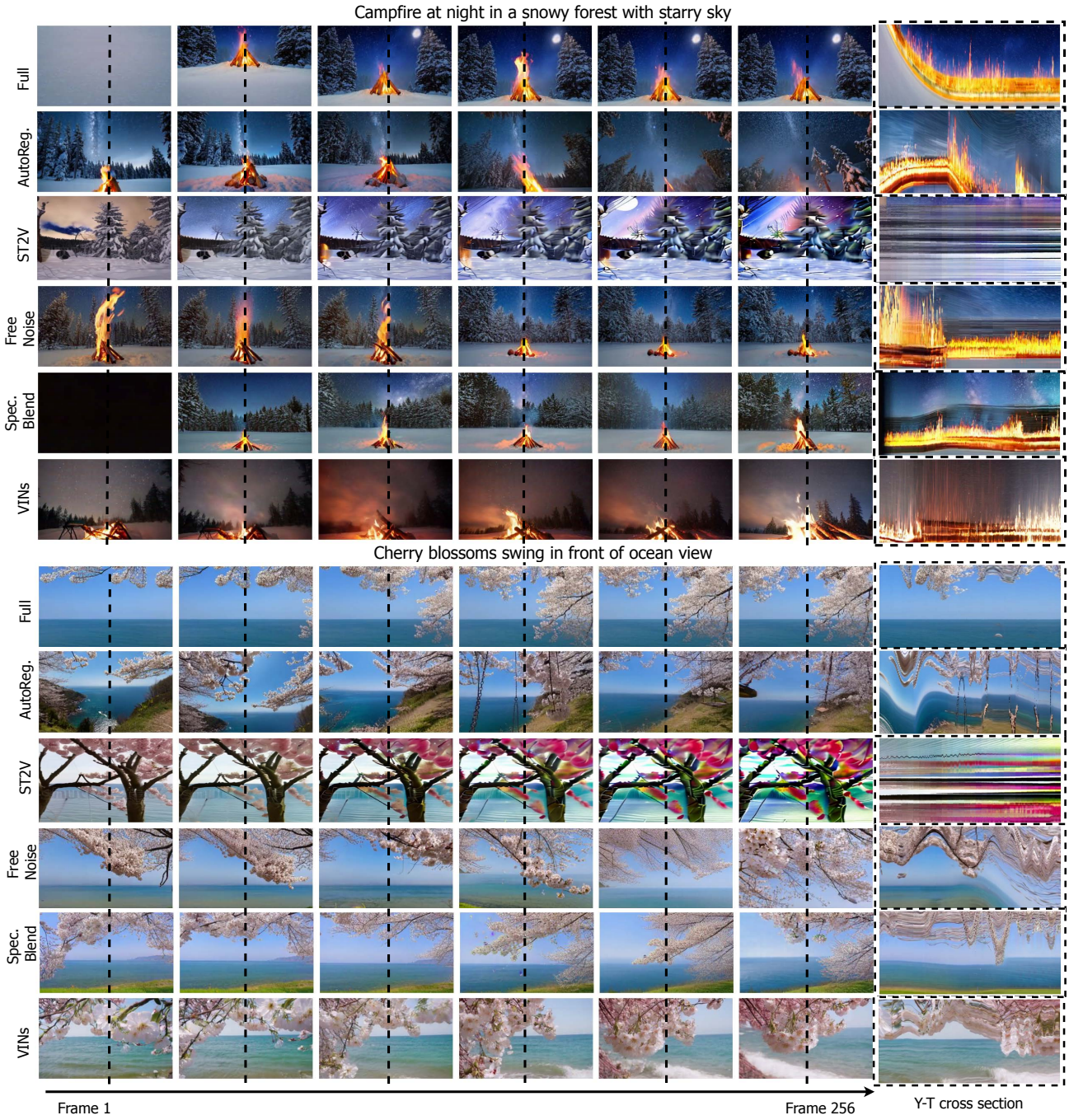

Figure 21. Qualitative visualizations across different methods.

11. Additonal Ablation Results

Ablation Experiments Setup. We used a smaller subset of prompts from the VBench suite for the ablation experiments. We extracted 25 prompts from the overall consistency prompts and, for each variant of the ablated model, generated two samples per prompt.

Global Token Ablation. Fig. 22 shows the effect of the global tokens. We ran the sampling chain under two settings: (1) global tokens and (2) without global tokens, initialized from the same noise. Identities of subjects in videos generated without global tokens tend to drift in subsequent frames.

12. Qualitative Interpretability

Figs. 23 and 24 show the VIN encoder attention maps on inputs. Fig. 23 displays the temporal dynamics of the attention patterns across videos. At the beginning of the video, attention is assigned to all objects within the scene. However, as the video progresses, we observe motion-aware encoding such that only the objects in motion are weighed significantly by the attention heads. For example, in the first example of Fig. 23, the attention heads focus on the man and the suitcase nearby and stay focused as the camera pans upward. As the camera stops and only the arms of the man move, attention from the suitcase is removed and transferred to the moving limbs. Similarly, in the third example, the attention pivots from the broader scene to the fingers and scissors in active motion while cutting the vegetables. Fig. 24 shows the object-centric attention over individual frames across different videos.

13. Training and Inference Algorithms for VINs

Algorithms 1 and 2 detail the training and inference algorithm, respectively, for the VIN-DiT coupling.

Algorithm 1 Training with VIN $\leftrightarrow$ DiT ensemble at t

Require: Noisy tokenized input X_t from $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Require: N_s chunk size, N_{local} local size
  $Z_t \leftarrow f_\alpha(X_t, t)$ ▷ Encode X_t via VINs
  for $i = 0, \cdots, \lfloor N/N_s \rfloor$ **do** ▷ Run in Parallel
    $X_t^i \leftarrow X_t[:, iN_s : (i+1)N_s]$
    $X_t^{local} \leftarrow X_t[:, iN_s - N_{local} : iN_s].detach()$
    $\hat{\epsilon}_t^i \leftarrow \epsilon_\theta(concat[X_t^{local},\ X_t^i,\ Z_t],\ t)$ ▷ Denoise
    $\hat{\epsilon}_t^i \leftarrow \hat{\epsilon}_t^i[:, N_{local} : N_{local} + N_s]$ ▷ Drop tokens
  end for
  $\hat{\epsilon}_t = concat\left[\hat{\epsilon}_t^0, \cdots, \hat{\epsilon}_t^{\lfloor N/N_s \rfloor}\right]$
  Gradient step on $\nabla_{\alpha,\theta} \left\| \hat{\epsilon}_t - \epsilon_t \right\|^2$

Algorithm 2 Inference with VIN $\leftrightarrow$ DiT ensemble

Require: Forward schedule α_t, Reverse Variance σ_t
Require: N_s chunk size, N_{local} local size, Timesteps T
  Set $x_T \sim \mathcal{N}(0, 1)$
  for $t = T-1, \cdots, 1$ **do**
    $Z_t \leftarrow f_\alpha(X_t, t)$
    for $i = 0, \cdots, \lfloor N/N_s \rfloor$ **do** ▷ Run in Parallel
      $X_t^i \leftarrow X_t[:, iN_s : (i+1)N_s]$
      $X_t^{local} \leftarrow X_t[:, iN_s - N_{local} : iN_s]$
      $\hat{\epsilon}_t^i \leftarrow \epsilon_\theta(concat[X_t^{local},\ X_t^i,\ Z_t],\ t)$
      $\hat{\epsilon}_t^i \leftarrow \hat{\epsilon}_t^i[:, : N_{local} + N_s]$
    end for
    if fuse tokens then $\hat{\epsilon}^t = \text{TokenFusion}(\hat{\epsilon}^t)$
    $\hat{\epsilon}_t = concat\left[\hat{\epsilon}_t^0, \cdots, \hat{\epsilon}_t^{\lfloor N/N_s \rfloor}\right]$
    $\mathbf{x}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_t\right) + \sigma_t \mathbf{z}$
  end for

14. User Study Design

We set up a comparison of VIN against other state-of-the-art methods considered in this work. A cohort of humans was presented with two 256-frame videos at a time, generated from the same prompt, and asked to rate which video was better with an option to choose that they prefer them equally. One of the videos was always generated using our method and the order of the videos was randomized. Raters were asked to assess the videos on two metrics: (1) overall appearance and (2) temporal consistency. See Fig. 25 for a screenshot of the survey page. We presented the user with comparisons over 45 prompts. The prompt was selected at random from a pool of 25 prompts. The 45 comparisons were equally divided as nine comparisons against each of the five methods. Overall, we received 100 assessments, each between VIN and other methods. For each metric under consideration, we reported the percentage with which VIN was deemed better (wins), comparable (draws), and worse (losses).

15. Test Prompts

We detail the prompt suite used across evaluations performed in this work.

15.1. Prompts used for VBench Evaluation

- Subject Consistency:
  1. A giraffe running to join a herd of its kind
  2. A car turning a corner
  3. A car accelerating to gain speed
  4. A train accelerating to gain speed
  5. A cat drinking water
  6. A dog enjoying a peaceful walk
  7. An airplane soaring through a clear blue sky

Figure 22. Qualitative visualizations with and without global tokens. Frames, where identity begins distorting, have been highlighted with a red box.

8. A cow running to join a herd of its kind
9. An airplane landing smoothly on a runway
10. A motorcycle accelerating to gain speed
11. A truck turning a corner

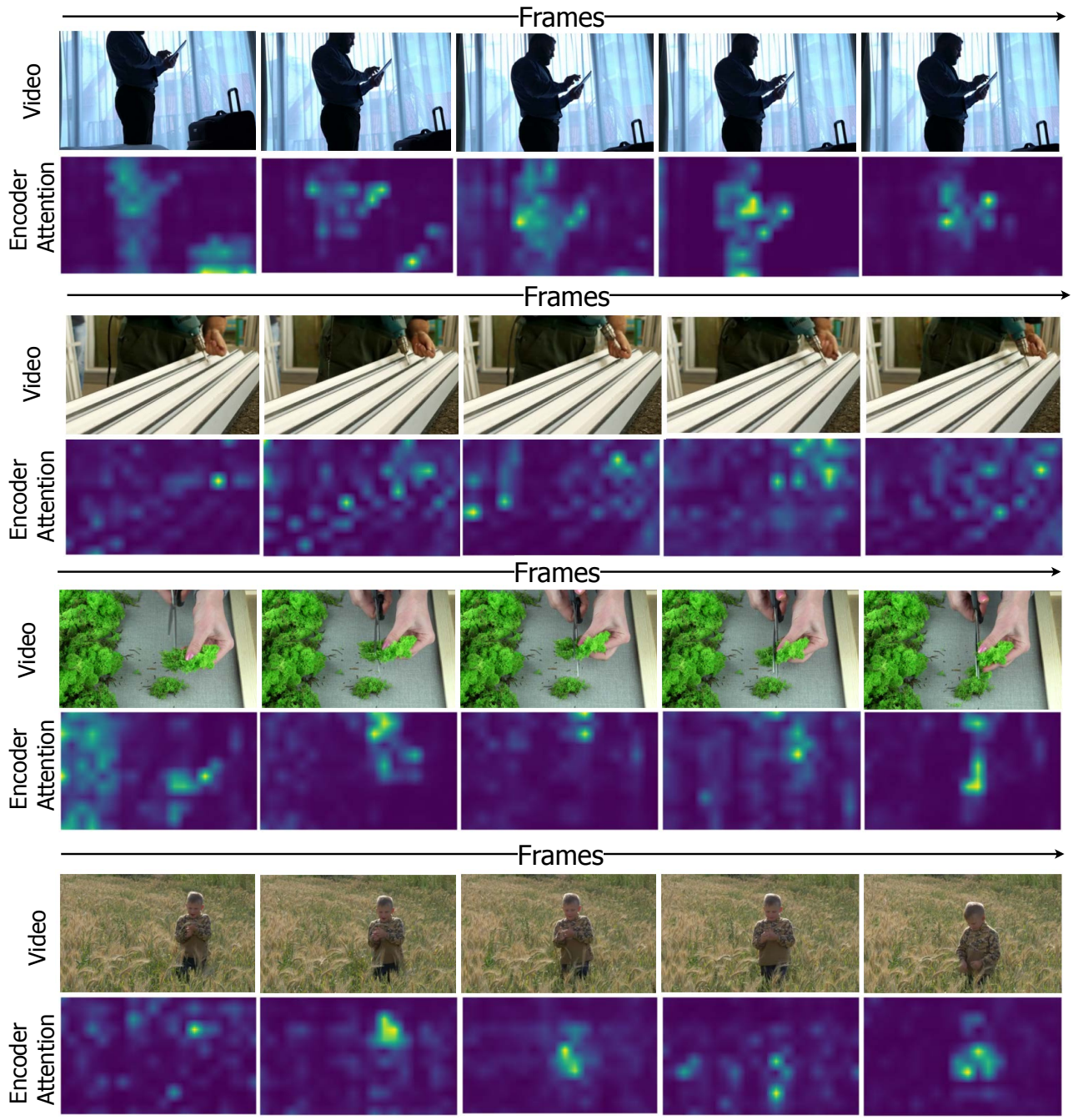

Figure 23. Qualitative visualizations across video and encoder attention. VIN exhibits motion-aware encoding, where it encodes dynamic objects as the video progresses.

12. A giraffe bending down to drink water from a river
13. A bicycle accelerating to gain speed
14. A car stuck in traffic during rush hour
15. A truck stuck in traffic during rush hour
16. An airplane accelerating to gain speed
17. A cat playing in park
18. A horse taking a peaceful walk
19. A cow chewing cud while resting in a tranquil barn
20. A dog running happily
21. A person drinking coffee in a cafe
22. A person walking in the snowstorm
23. A zebra bending down to drink water from a river

Figure 24. Qualitative visualizations across frames and encoder attention.

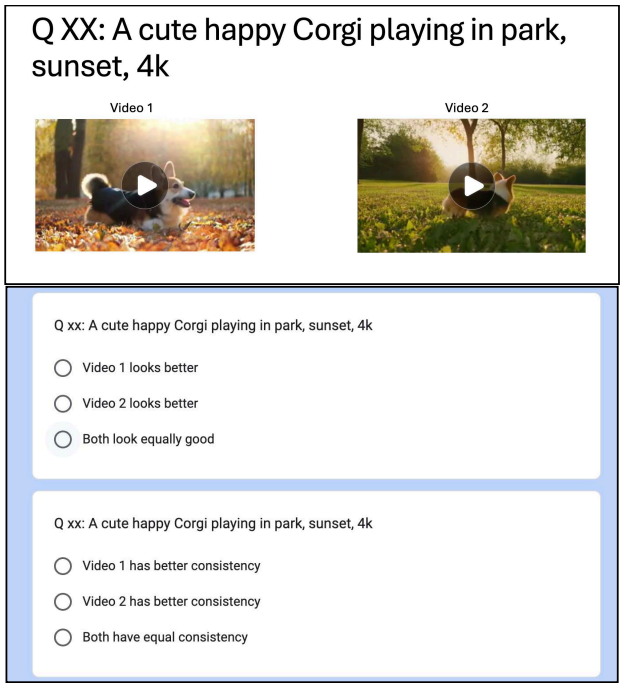

Figure 25. User study design

24. A bicycle leaning against a tree
25. A cow bending down to drink water from a river

- Background Consistency:
  1. Underwater coral reef
  2. Phone booth
  3. Hospital
  4. Arch
  5. Glacier
  6. Jail cell
  7. Sky
  8. Highway
  9. Classroom
  10. Basement
  11. Staircase
  12. Bathroom
  13. Volcano
  14. Construction site
  15. Valley
  16. Beach
  17. Ballroom
  18. Fountain
  19. Skyscraper
  20. Raceway
  21. Office
  22. Ski slope
  23. Golf course
  24. Tower
  25. Cliff
- Motion Smoothness
  1. A person washing the dishes
  2. A person giving a presentation to a room full of colleagues
  3. A sheep bending down to drink water from a river
  4. A horse taking a peaceful walk
  5. A bus stuck in traffic during rush hour

6. A bicycle accelerating to gain speed
7. A car stuck in traffic during rush hour
8. A truck turning a corner
9. A dog drinking water
10. A cat playing in park
11. A motorcycle accelerating to gain speed
12. A dog running happily
13. A bird building a nest from twigs and leaves
14. A person swimming in ocean
15. A giraffe taking a peaceful walk
16. A cow chewing cud while resting in a tranquil barn
17. A person playing guitar
18. A train crossing over a tall bridge
19. A truck slowing down to stop
20. A train speeding down the tracks
21. A car turning a corner
22. A zebra running to join a herd of its kind
23. A bird soaring gracefully in the sky
24. A motorcycle cruising along a coastal highway
25. A truck accelerating to gain speed

- Temporal Flickering
  1. A tranquil tableau of a bunch of grapes
  2. A tranquil tableau of kitchen
  3. A tranquil tableau of palace
  4. A tranquil tableau in the heart of Plaka, the neoclassical architecture of the old city harmonizes with the ancient ruins
  5. In a still frame, parking lot
  6. A toilet, frozen in time
  7. In a still frame, a tranquil pond was fringed by weeping cherry trees, their blossoms drifting lazily onto the glassy surface
  8. A tranquil tableau of a chair
  9. A tranquil tableau of the jail cell was small and dimly lit, with cold, steel bars
  10. A tranquil tableau of an antique bowl
  11. A tranquil tableau at the edge of the Arabian Desert, the ancient city of Petra beckoned with its enigmatic rock-carved façades
  12. A laptop, frozen in time
  13. A tranquil tableau of a beautiful wrought-iron bench surrounded by blooming flowers
  14. A tranquil tableau of a wooden bench in the park
  15. In a still frame, in the vast desert, an oasis nestled among dunes featuring tall palm trees and an air of serenity
  16. A tranquil tableau of a bowl on the kitchen counter
  17. A tranquil tableau of a country estate's library featured elegant wooden shelves
  18. In a still frame, a tranquil Japanese tea ceremony room, with tatami mats, a delicate tea set, and a bonsai tree in the corner
  19. Indoor gymnasium, frozen in time
  20. Static view on a desert scene with an oasis, palm trees, and a clear, calm pool of water
  21. A tranquil tableau of restaurant
  22. A tranquil tableau of a dining table
  23. A tranquil tableau of a tranquil lakeside cabin nestled among tall pines, its reflection mirrored perfectly in the calm water
  24. In a still frame, nestled in the Zen garden, a rustic teahouse featured tatami seating and a traditional charcoal brazier
  25. A tranquil tableau of barn
- Temporal Style
  1. A boat sailing leisurely along the Seine River with the Eiffel Tower in background, tilt down
  2. A boat sailing leisurely along the Seine River with the Eiffel Tower in background, zoom in
  3. A couple in formal evening wear going home get caught in a heavy downpour with umbrellas, tilt up
  4. A boat sailing leisurely along the Seine River with the Eiffel Tower in background, tilt up
  5. Snow rocky mountains peaks canyon. Snow blanketed rocky mountains surround and shadow deep canyons. The canyons twist and bend through the high elevated mountain peaks, with an intense shaking effect
  6. The bund Shanghai, in super slow motion
  7. A boat sailing leisurely along the Seine River with the Eiffel Tower in background, featuring a steady and smooth perspective
  8. The bund Shanghai, zoom in
  9. A couple in formal evening wear going home get caught in a heavy downpour with umbrellas, pan left
  10. A boat sailing leisurely along the Seine River with the Eiffel Tower in background, pan right
  11. A shark is swimming in the ocean, featuring a steady and smooth perspective
  12. A couple in formal evening wear going home get caught in a heavy downpour with umbrellas, zoom out
  13. A couple in formal evening wear going home get caught in a heavy downpour with umbrellas, with an intense shaking effect
  14. An astronaut flying in space, with an intense shaking effect
  15. An astronaut flying in space, in super slow motion
  16. A shark is swimming in the ocean, tilt up
  17. An astronaut flying in space, racking focus
  18. The bund Shanghai, pan right
  19. Gwen Stacy reading a book, tilt up
  20. The bund Shanghai, racking focus
  21. A shark is swimming in the ocean, pan right
  22. A cute happy Corgi playing in park, sunset, tilt up
  23. A cute happy Corgi playing in park, sunset, pan right
  24. A cute happy Corgi playing in park, sunset, featuring a steady and smooth perspective

25. A couple in formal evening wear going home get caught in a heavy downpour with umbrellas, in super slow motion

- Overall Consistency
  1. A beautiful coastal beach in spring, waves lapping on sand by Hokusai, in the style of Ukiyo
  2. A beautiful coastal beach in spring, waves lapping on sand by Vincent van Gogh
  3. A car moving slowly on an empty street, rainy evening
  4. A drone flying over a snowy forest
  5. A drone view of celebration with Christmas tree and fireworks, starry sky - background
  6. A panda playing on a swing set
  7. A panda standing on a surfboard in the ocean in sunset
  8. A space shuttle launching into orbit, with flames and smoke billowing out from the engines
  9. A teddy bear washing the dishes
  10. An artist brush painting on a canvas close up
  11. An astronaut feeding ducks on a sunny afternoon, reflection from the water
  12. An astronaut flying in space
  13. An ice cream is melting on the table
  14. An oil painting of a couple in formal evening wear going home get caught in a heavy downpour with umbrellas
  15. Few big purple plums rotating on the turntable. Water drops appear on the skin during rotation. Isolated on the white background. Close-up. Macro
  16. Golden fish swimming in the water
  17. Happy dog wearing a yellow turtleneck, studio, portrait, facing camera, dark background
  18. Motion colour drop in water, ink swirling in water, colourful ink in water, abstraction fancy dream cloud of ink
  19. Sewing machine, old sewing machine working
  20. Time lapse of sunrise on Mars
  21. Turtle swimming in ocean
  22. Two pandas discussing an academic paper
  23. Yellow flowers swinging in the wild
  24. Yoda playing guitar on the stage

15.2. Prompts used for User Study

1. A drone flying over a snowy forest
2. A panda standing on a surfboard
3. A teddy bear washing dishes
4. Happy dog wearing a yellow turtleneck, studio, portrait, facing camera, dark background
5. Golden fish swimming in the ocean
6. Yellow flowers swing in the wind
7. Two pandas discussing an academic paper
8. Smiling elderly gentlemen with rimmed glasses, tweed jacket, studio, standing still, burgundy background
9. Bear trying to learn calculus
10. Timelapse of sunrise on Mars
11. A giraffe running to join a herd of its kind
12. Cherry blossoms swinging by the ocean
13. Campfire at night in a snowy forest with starry sky in the background
14. A dog swimming
15. Beer pouring into a glass low-angle, wide shot
16. A raccoon wearing a suit playing the trumpet
17. A cat wearing sunglasses and working as a lifeguard at a pool
18. A fat rabbit wearing a purple robe walking through a fantasy landscape
19. Back view on young woman dressed in a yellow jacket walking in the forest
20. A shark swimming in clear Carribean ocean
21. A petri dish with a bamboo forest growing within it that has tiny red pandas running around
22. A swarm of bees flying around their hive
23. A fantasy landscape
24. Aerial view of a snow-covered mountain
25. A dog wearing a Superhero outfit with red cape flying through the sky