Diverse Video Generation with Determinantal Point Process-Guided Policy Optimization

Supplementary Material

Table of Contents

1. Scalability and Diversity Trade-offs.

Our method is model-agnostic and scales effectively to long-form generation, as demonstrated in **Fig. b** in 1, where it produces diverse 1-minute videos using the Self-Forcing++ model. However, this enhancement involves a configurable trade-off; as shown in the ablation study in **Fig. a** in 1, intentionally increasing the diversity weight λ_{div} can lead to "excessive diversity". In cases with underspecified prompts, such as "a dog," a very high λ_{div} may cause the policy to introduce extra attributes to maximize diversity, drifting from a simpler depiction while still technically satisfying the prompt.

2. Video-Level Reward.

While our framework operates in prompt space for efficiency and plug-and-play deployment, it naturally supports video-level feedback. As shown in Tab. 2, incorporating a video reward model (VRM) via Long-RL yields further gains over our base method, achieving the highest TCE (25.88) and TIE (42.53), demonstrating that richer reward signals translate directly into improved diversity.

Table 1. MSR-VTT Results

Method	TCE ↑	TIE ↑	Vendi ↑	CLIP ↑
Wan	13.02	23.40	10.50	0.291
Ours	**16.37**	**27.85**	**12.89**	**0.295**

Table 2. Comparison of diversity and alignment metrics across methods. Higher is better (↑) for all metrics.

Method	TCE ↑	TIE ↑	Vendi ↑	CLIP ↑
Ours (VRM)	25.88	42.53	13.38	0.303
Ours	23.79	41.26	12.44	0.327
Wan	12.35	23.27	9.70	0.324

3. Generalization.

Tab. 1 reports results on MSR-VTT, where our method consistently outperforms Wan 2.1 across all metrics (TCE: 13.02→16.37, TIE: 23.40→27.85, VENDI: 10.50→12.89, CLIP: 0.291→0.295), confirming that DPP-GRPO generalizes beyond the training distribution.

4. Additional Qualitative Examples

Please refer to the end of supplementary webpage (.html file) or the supplementary folder for all the videos in our main paper, as well as additional videos.

$(\lambda_{\text{div}}, \lambda_{\text{rel}})$	TCE	TIE	CLIP
(0.9, 0.1)	16.910	24.256	0.285
(0.5, 0.5)	16.709	23.945	0.302
(0.1, 0.9)	16.002	23.735	0.305

Table 3. Ablation on reward weights $(\lambda_{\text{div}}, \lambda_{\text{rel}})$. We vary the balance between diversity and relevance during training.

5. Video Prompts

Due to space limitations, we could not include the full prompts in the main paper. However, we provide the complete prompts for several videos generated by our method as well as and system prompt used in tables 8, 7, 12, 14, 11, 9, 13, 10, and 5.

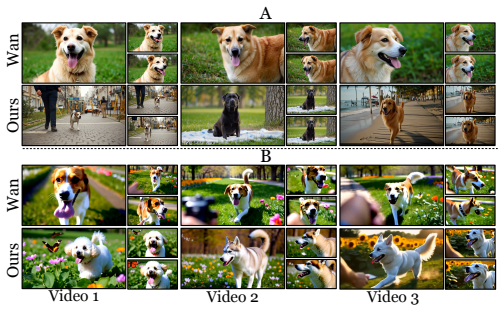

Figure 1. (a) **Excessive Diversity**: Increasing the diversity weight λ_{div} for underspecified prompts (e.g., "a dog") can cause the policy to introduce extra attributes to maximize diversity. (b) **Long-video Scalability**: Our model-agnostic approach scales to 1-minute video generation using the Self-Forcing++ model.

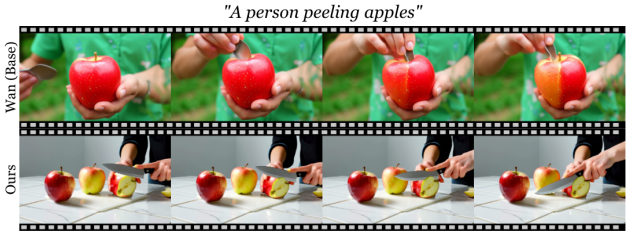

Figure 2. An example failure case of our method where the temporal video quality depends on the base model's ability.

6. User Study Details

We provide a screenshot of our user study in Fig. 4. Users are shown 4 videos generated by a given method, and asked to rate the diversity and text alignment on a Likert scale 1-5.

7. Failure Cases

As shown in Figure 2, when the base model struggles with fine-grained motions in complex actions (e.g., peeling actions), our method also struggles in overcoming this constraint, as temporal dynamics are governed by the model's temporal attention values rather than prior conditioning.

8. Additional Ablations

Set Size	TCE	TIE	CLIP
2	15.131	22.939	0.308
5	16.570	23.936	0.308
8	16.959	24.464	0.306
10	15.546	23.040	0.306

Table 4. Ablation on reference set size $|\mathcal{R}_q|$. Small multi-reference sets (5–8 examples) yield the best diversity, while larger sets show diminishing returns.

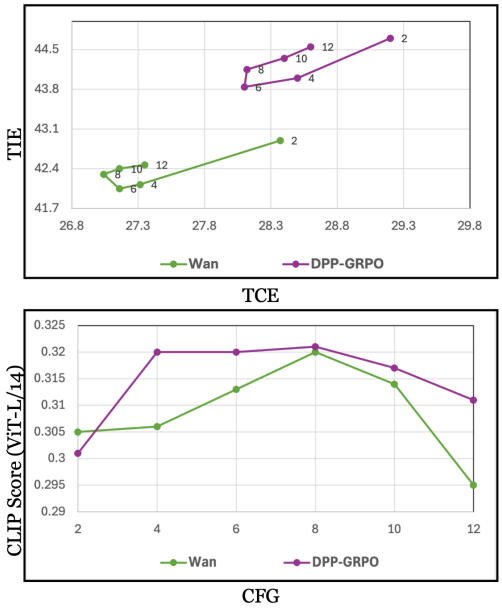

Figure 3. **CFG ablation.** Diversity (TCE/TIE) and fidelity (CLIP) across different CFG values for Wan and our DPP–GRPO model.

8.1. Effect of Reference Set Size during Training

We study how the size of the reference set $|\mathcal{R}_q|$ influences training dynamics and diversity. Table 4 shows that using a small multi-reference set (5-8 examples) yields the higher improvements in TCE and TIE, confirming that exposing the policy to several video-grounded modes produces a more reliable marginal-gain signal. Performance degrades when the set becomes too large ($|\mathcal{R}_q| = 10$), consistent with the diminishing-returns property of the DPP log-determinant and the increased variance of similarity estimates in larger matrices. CLIP alignment remains stable across settings, with only a mild drop for larger sets. Overall, a modest reference set (5-8 samples) provides the best balance between diversity gain and semantic stability.

8.2. Ablation on CFG Scale

Figure 3 presents an ablation study measuring the effect of classifier-free guidance (CFG) values on both diversity and fidelity for the baseline Wan model and our `DPP-GRPO` method. We sweep the CFG scale over $\{2, 4, 6, 8, 10, 12\}$ and report (1) the joint behavior of TCE and TIE, and (2) CLIP alignment scores. The upper plot in Figure 3 shows that our method consistently achieves higher TCE and TIE across all CFG values. While both models exhibit non-monotonic trajectories as CFG increases, the `DPP-GRPO`frontier shifts upward in the TCE-TIE plane, demonstrating that our objective is robust to abrupt CFG adjustments. This indicates that our method remains effective under a wide range of decoding hyperparameters and does not depend on care-

fully tuned CFG settings. The lower plot in Figure 3 reports CLIP similarity scores. Both models exhibit the expected behavior in which CLIP scores peak at moderate guidance scales (CFG $\approx$ 6-8). Importantly, our method maintains fidelity comparable to the baseline across all settings, and achieves the highest CLIP score at CFG $= 6$.

8.3. Ablation on λ hyperparameters

Table 3 shows how varying the reward weights $(\lambda, \lambda_{\text{div}})$ affects performance. Increasing the diversity weight ($\lambda_{\text{div}} = 0.9$) yields the highest TIE/TCE, indicating stronger diversity, while moderately CLIP alignment. Conversely, prioritizing relevance ($\lambda_{\text{rel}} = 0.9$) improves CLIP but reduces diversity. The balanced setting $(0.5, 0.5)$ provides a middle ground across all metrics. Overall, the ablation highlights the expected trade-off: higher diversity weight improves motion and variation, while higher relevance weight preserves semantic fidelity.

9. Dataset Generation

Our dataset is constructed in two stages. First, we extend the VBench prompt categories by using chain-of-thought prompting to generate approximately 350 prompts for each of the 7 VBench categories. We additionally include the original VBench prompts as part of our evaluation set. This ensures that our prompt space remains grounded in a widely used video-generation benchmark. In the second stage, each curated base prompt is expanded into 10 diverse variants using an iterative, two-agent reasoning framework. An architect agent proposes candidate expansions (system prompt shown in Table 5), and a critic agent evaluates the videos produced from these expansions using established video-level metrics: TCE/TIE for temporal diversity, CLIP for semantic alignment, and VideoScore for perceptual quality. Only expansions that satisfy diversity and alignment criteria are retained. This procedure ensures that the final dataset is not only textually diverse but also grounded in video-level behavior.

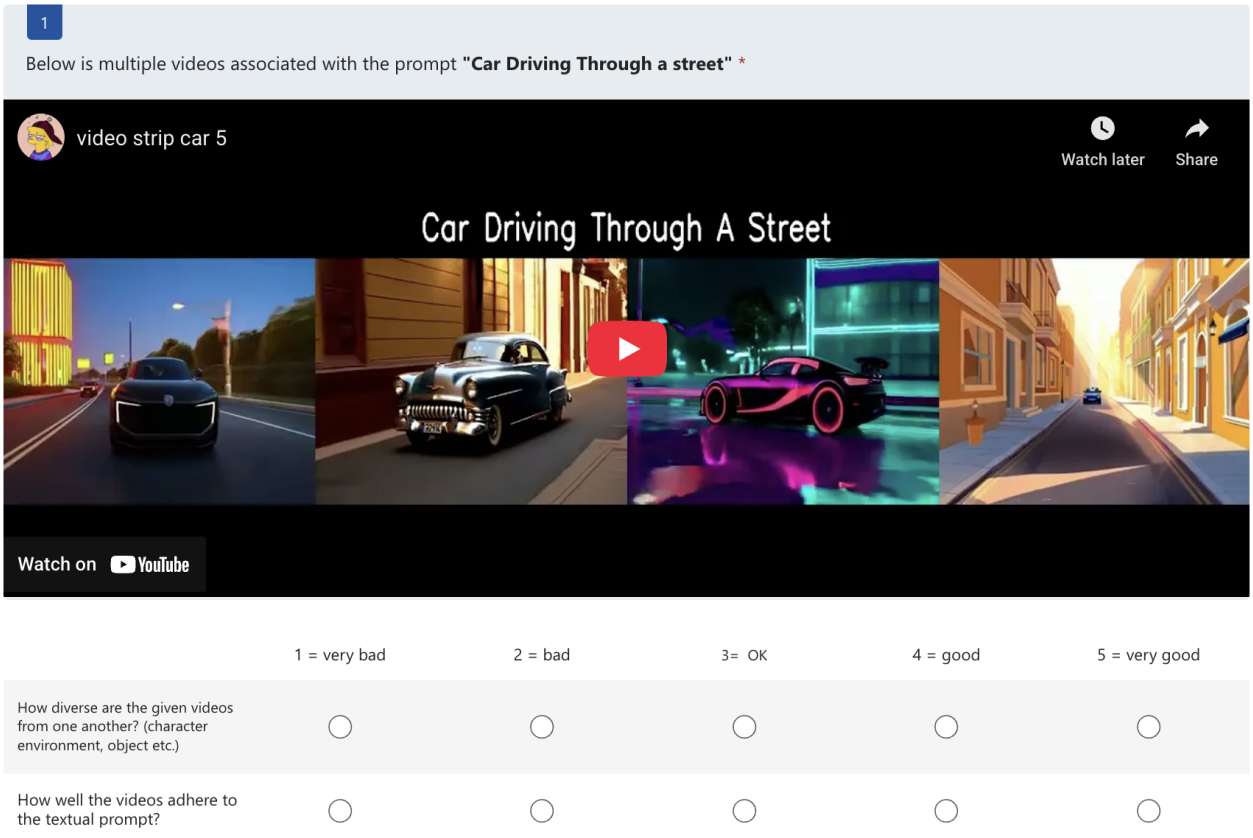

Figure 4. A screenshot of our user study where 4 videos from a given method are shown to the users.

Table 5. System Prompt

System Prompt Content

```
    You are an expert at prompt expansion.  Given a base prompt, expand it into at most four sentences.
       1.  Preserve the main object(s), number, and action (no semantic drift).
       2.  Expand along dimensions:  environment, object attributes, subject variation, time/season,
perspective, narrative/action, character appearance, or cultural context.
       4.  Use English only.
       Example:
       Base prompt:  a cat sitting on a windowsill.
       Good expansion:  "A fluffy Siamese cat with bright blue eyes sits on a wooden windowsill,
watching the rainy evening street in a warm painterly style."
       Bad expansion:  "A group of cats running through a sunny garden, chasing butterflies." (changes
number, action, setting).
    You will be given one user query.
    Your task:  write a new expansion that is diverse from the references while still preserving
meaning.
```

Table 6. Video Prompts

Water Lily rests on a calm pond

Video 1: "A minimalist depiction of a white water lily resting at the edge of a calm pond, bold outlines, restrained greens, and a crisp lines."
Video 2: "A soft watercolor scene showing a water lily resting on a calm pond, the bloom painted in gentle pink tones with a soft wash and a gentle backdrop of water reflections."
Video 3: "A top-down, aerial view of a water lily resting on a calm pond, the outline tracing a perfect circle while the pool's mirror-like surface reflects the scene in photorealistic detail."
Video 4: "A minimalist vector style renders a water lily resting at the edge of a calm pond, the lotus represented as a clean silhouette against a calm, mirror-like surface with subtle radial symmetry."

Table 7. Video Prompts

A cat eating food out of a bowl on a sidewalk

Video 1: "A shot of a cat on a sunlit sidewalk eating from a shallow ceramic bowl, food a bright color, street life and people blurred in the background."
Video 2: "In a street cafe atmosphere, a Siamese cat eats from a ceramic bowl on a sunlit sidewalk, the bowl catching the glow and the cat's fur shimmering with the passing traffic in soft watercolor."
Video 3: "In a sunlit courtyard, a fluffy Persian cat sits on a bright cobblestone, eating kibble from a ceramic bowl while the sun casts a warm glow across the scene."
Video 4: "From a ground-level perspective, a cat in a bold patterned jacket eats from a bright ceramic bowl on a sunlit city sidewalk, the bowl's rim catching a glint of sun while the background bustles with activity"

Table 8. Video Prompts

A boat gliding across a lake at twilight

Video 1: "A moody urban scene: a sleek glass-bottomed ferry glides across a glittering lake at twilight, the city skyline glowing behind and silhouettes etched in the glass, captured in digital painting with a dramatic color gradient."
Video 2: "A cinematic time-lapse across a lake at twilight showing a boat gliding across the water, the sky's changing light refracting on the surface and the water reflecting a pale, dreamlike color."
Video 3: "Across a lake, a boat glides at twilight; the camera tracks the craft from a low-angle perspective on a grassy bank, the water shifting from silver to pink and the horizon turning a deep indigo."
Video 4: "A hyperreal CGI rendering of a sleek boat gliding across a glassy lake at twilight, reflective water, starry skies, and cool blue-green hues to capture the moment."

Table 9. Video Prompts

A giraffe bending to sip water from a sunlit savanna pool

Video 1: "From a low-angle view, a giraffe bends to sip water from a sunlit savannah pool, long eyelashes brushing the surface and the sunlit grasses shimmering in the background"
Video 2: "A painterly color-graded shot with warm sunset hues shows a giraffe bending to sip water from a sunlit savanna pool, distant grasses glow and a subtle horizon line guides the eye.",
Video 3: "A painterly color-graded frame turning the sunlit savanna into amber tones, a giraffe bending to sip water from a sunlit savanna pool, rendered in digital gouache with thick brushstrokes.",
Video 4: "A photorealistic close-up of a giraffe bending to sip water from a sunlit savanna pool, as the sunbeams split the grasses and the water shimmers with color."

Table 10. Video Prompts

A skateboarder performs jumps

Video 1: " "On a sunlit park plaza, a teenage Black girl with vibrant hair performs a set of jumps and grinds on a street-style board, ground reflects light."
Video 2: "On a concrete skatepark plaza at sunset, a Black teenage skateboarder performs a series of jumps from a deck, the crowd watching as the skateboarder leaps with confidence, captured in cinematic shot."

Table 11. Video Prompts

A fantasy landscape

Video 1:"A panorama-style shot of a forest glade at twilight, holographic flowers and glow moss lighting the scene as a breeze moves the leaves."
Video 2: "A surreal yet coherent dreamlike landscape where frost-draped trees shimmer under aurora, the moonlight catching tiny frost crystals, and a faint, otherworldly air pervades the scene."
Video 3: "In a world where magic meets science, a sun-dappled desert rises around a mysterious castle; crystals glow with luminosity as water droplets create pools across the scene."
Video 4: "A dawn panorama across a fantastical valley with a river, a shimmering dragonfly-like wing across the sky, and the air thick with mist, rendered in pastel watercolor with a warm glow."

Table 12. Video Prompts

A coffee cup sitting on a wooden table

Video 1:"A cinematic top-down view in a modern kitchen shows a porcelain coffee cup on a wooden table, a single grain visible in the wood and a single drop of condensation forming on the rim."
Video 2: "An artisanal coffee mug with a rustic ceramic rim rests on a wooden table, the scene rendered in warm watercolor with soft brushstrokes and a delicate steam curl above the mug"
Video 3: "A studio macro on a glass coffee cup on a wooden table, the cup resting in the center while the wooden grain and the steam in the air form a high-contrast backdrop.",
Video 4: "A vintage-inspired setting shows a glass coffee cup on a wooden table, sunbeams filtering through lace curtains and dust motes dancing in the air as the cup glows softly."

Table 13. Video Prompts

A person kayaking or canoeing

Video 1:"A South Asian woman in a wetsuit paddles a bright pink kayak across a tranquil river, sunlight turning the water to copper and casting a warm halo around the boat."
Video 2: "A dawn scene where a person paddles a kayak toward a misty marsh; soft washes of pink and teal, with a shallow, warm glow on the water."
Video 3: "A sunlit meadow scene where a person in a neon-green paddling jacket is canoeing along a quiet river, wildflowers blooming nearby and the sun casting a warm glow across the water."
Video 4: "A person in a kayak glides along a tranquil river at dawn, sun rising and the water shimmering with golds, rendered in dreamy mood with loose brushwork and soft edges."

Table 14. Video Prompts

A dog playing ball on the beach

Video 1: In a sunset beach scene, a beagle with a bright coat and long ears romps after a ball on the sand, waves lapping and a light breeze curling the hair, watercolor mood."
Video 2: "A cinematic color-graded shot with vibrant warm hues and cool shadows of a golden retriever at foggy weather, playing with a ball on a sunlit beach, high-angle view from above."

Table 15. Video Prompts

A fox walking through a forest

Video 1:"A photorealistic dawn scene of a red fox walking through a forest, warm light filtering through a canopy of birches and needles, dewy moss and a distant stream catching the early sun."
Video 2:"A white fox walking through a sunlit forest, soft brushwork greens and browns; the scene is dreamlike with a slight hint of surrealism."
Video 3: "A surreal concept envisions a fox walking through a dreamlike forest, trees bending to form a path and light washing the fur in soft tones."
Video 4: "A hyper-realistic CGI depiction of a fox walking through a forest at dawn, thick fur rendered with photoreal texture, golden hour light highlighting the forest floor."

Table 16. Video Prompts

A cyclist riding along a lakeside trail

Video 1:"A sunset sequence shows a cyclist riding along a lakeside trail, the sun casting warm light on the rider and the water reflecting hues of pink."
Video 2:"A vintage style rendering of a cyclist riding along a lakeside trail, bold shapes and warm sunset hues, retro typography for the route and destination."
Video 3: "A painterly color-graded frame turning the greens to emerald and the sky to a pale cerulean, as the cyclist glides along a lakeside trail, water reflections shimmering in the frame"
Video 4: "A cyclist on a lakeside trail glides toward the camera, reflections on the water shimmering and ripples spreading as a tranquil backdrop renders the scene cinematic."

Table 17. Video Prompts

A lantern swaying softly on windy night

Video 1:"A painterly color-graded scene where the lantern shifts to warm amber, the night breeze makes the light flutter and the background blurs into a dreamlike wash."
Video 2:"A painterly night scene showing a lantern swaying softly on a windy night in a quiet street, the glow diffusing through the wind and casting a warm halo across the cobblestones"
Video 3: "Isometric landscape vignette: a lantern floats in a serene park at twilight; ornate, graphic lanterns define natural shapes in geometric forms; soft pastel sky overhead adds a cozy, calm mood."
Video 4: "A cinematic long-shot with a lantern swaying softly on a windy night, a coastal boardwalk as the backdrop, mist rolling and the lantern creating a warm halo around the walker."