EIDT-V: Exploiting Intersections in Diffusion Trajectories for Model-Agnostic, Zero-Shot, Training-Free Text-to-Video Generation

Supplementary Material

9. Test Prompts

This section details the 50 prompts used to generate videos for our evaluation, along with the rationale behind their selection. The prompts were carefully designed to span a wide range of scenarios, including natural phenomena, object transformations, motion dynamics, and creative interpretations. This diversity ensures a comprehensive assessment of the model's capabilities across various aspects of text-to-video generation.

1. **A flower blooming from a bud to full bloom over time.** *Rationale:* Evaluates the model's ability to depict time-lapse growth with smooth transitions.
2. **A cat chasing a laser pointer dot across the room.** *Rationale:* Tests motion tracking and dynamic object interactions.
3. **A rotating 3D cube changing colors.** *Rationale:* Assesses rendering of 3D rotation and color transitions.
4. **A sunrise over the mountains turning into daytime.** *Rationale:* Evaluates depiction of natural phenomena and transitions in lighting conditions.
5. **A person morphing into a wolf under a full moon.** *Rationale:* Challenges the model's ability to handle complex transformations and creative scenarios.
6. **Raindrops falling into a puddle creating ripples.** *Rationale:* Tests fluid dynamics rendering and subtle animation effects.
7. **A city skyline transitioning from day to night with lights turning on.** *Rationale:* Evaluates handling of complex lighting transitions in urban scenes.
8. **An apple falling from a tree and bouncing on the ground.** *Rationale:* Assesses motion physics and interactions with gravity.
9. **A hand drawing a circle on a whiteboard.** *Rationale:* Tests precision in hand movements and sequential drawing actions.
10. **An ice cube melting into water.** *Rationale:* Evaluates the depiction of state changes from solid to liquid.
11. **A rocket launching into space and disappearing into the stars.** *Rationale:* Tests sequential events and scale changes in dynamic scenarios.
12. **A chameleon changing colors on a branch.** *Rationale:* Challenges the model's ability to handle color transitions and blending with surroundings.
13. **A balloon inflating and then popping.** *Rationale:* Evaluates expansion dynamics and sudden transitions.
14. **A paper airplane flying across a classroom.** *Rationale:* Tests object motion within a setting and interactions with the environment.
15. **Clouds forming and then dissipating in the sky.** *Rationale:* Assesses rendering of natural elements and gradual changes.
16. **A cup of coffee being poured with steam rising.** *Rationale:* Tests liquid dynamics and fine details like steam.
17. **A clock's hands moving fast-forward from noon to midnight.** *Rationale:* Evaluates representation of time passage and object motion.
18. **A caterpillar transforming into a butterfly.** *Rationale:* Tests depiction of life cycles and metamorphosis.
19. **A book opening and pages flipping.** *Rationale:* Evaluates detailed object movements and sequential actions.
20. **A snowman melting under the sun.** *Rationale:* Tests weather effects and melting animations.
21. **A traffic light cycling from red to green.** *Rationale:* Evaluates color changes and timing sequences.
22. **A fish jumping out of water and diving back in.** *Rationale:* Tests motion through different mediums and splash effects.
23. **An artist painting a canvas with a brush.** *Rationale:* Assesses fine motor actions and the process of creation.
24. **A spinning globe showing continents passing by.** *Rationale:* Tests rotational motion and geographical accuracy.
25. **Leaves falling from a tree in autumn.** *Rationale:* Evaluates natural motions and seasonal transitions.
26. **A car transforming into a robot.** *Rationale:* Challenges the model with complex object transformations.
27. **A candle burning down with the flame flickering.** *Rationale:* Tests gradual reduction and subtle lighting effects.
28. **A soccer ball being kicked into a goal.** *Rationale:* Assesses motion, action sequences, and interactions.
29. **A river flowing through a forest.** *Rationale:* Evaluates fluid motion and natural scenery rendering.
30. **A rainbow appearing after rain.** *Rationale:* Tests depiction of weather transitions and color spectrum rendering.
31. **An eclipse where the moon passes in front of the sun.** *Rationale:* Assesses celestial motion and lighting effects.
32. **A horse galloping across a field.** *Rationale:* Tests animal motion and interaction with natural environments.
33. **Popcorn popping in a microwave.** *Rationale:* Evaluates rapid, random movements and cooking processes.
34. **A kaleidoscope pattern changing shapes and colors.**

Rationale: Tests abstract patterns and continuous transformations.

35. **A glass shattering into pieces when dropped.** *Rationale:* Challenges the model with sudden fragmentation and physics.
36. **An astronaut floating in space waving.** *Rationale:* Tests human figures and movement in zero gravity.
37. **A plant growing from a seed to a sapling.** *Rationale:* Evaluates depiction of growth over time.
38. **Fireworks exploding in the night sky.** *Rationale:* Tests bright, dynamic visuals in a dark setting.
39. **A dog wagging its tail happily.** *Rationale:* Assesses animal emotions and natural movements.
40. **A compass needle spinning and settling north.** *Rationale:* Tests rotational motion and stabilization dynamics.
41. **An umbrella opening up during rainfall.** *Rationale:* Evaluates object transformations and interactions with weather.
42. **A stop-motion animation of clay figures moving.** *Rationale:* Tests frame-by-frame animation styles.
43. **A battery draining from full to empty.** *Rationale:* Assesses gradual representation of depletion over time.
44. **A puzzle being assembled piece by piece.** *Rationale:* Evaluates sequential object placement and completion.
45. **A windmill's blades turning in the breeze.** *Rationale:* Tests rotational motion influenced by wind.
46. **A snake slithering through the grass.** *Rationale:* Assesses complex body movements in a natural setting.
47. **A paintbrush changing colors as it moves.** *Rationale:* Tests motion-linked color transitions.
48. **A volcano erupting with lava flowing.** *Rationale:* Evaluates dynamic natural events and fluid motion.
49. **An eye blinking slowly.** *Rationale:* Tests subtle facial movements and precise timing.
50. **A paper crumpling into a ball.** *Rationale:* Challenges the model with complex folding and texture changes.

These prompts ensure a diverse evaluation of model capabilities, covering natural phenomena, motion dynamics, and creative transformations.

10. Hyperparameter Selection

Hyperparameter tuning was a critical step in optimizing the performance of our video generation models, particularly with respect to temporal coherence, visual fidelity, and prompt adherence. We conducted a systematic grid search for SDXL and performed manual tuning for SD1.5 and SD3 to identify the most effective configurations for each model.

10.1. Hyperparameters Considered

The following key hyperparameters were explored during the grid search:

- **Batch Size:** Values of 1, 2, and 3 were tested to balance GPU memory usage and frame coherence. Larger batch sizes can improve smoothness across frames by enabling better context preservation but increase memory requirements.
- **Intersection Strategy:** To ensure temporal continuity between frames, two strategies were compared:
  - **First:** Each frame intersects with a static base image (batch size = num frames - 1).
  - **Previous:** Each frame intersects with the last frame from the previous batch.
- **Guidance Scale:** A range of values from 3.0 to 13.0 was tested to balance adherence to text prompts against visual diversity. Higher values generally emphasize prompt alignment but may reduce variability.
- **Multi-Prompt Strategy:** For models supporting multiple text inputs, we evaluated different strategies:
  - **PreviousFrame:** Using the text of the previous frame as secondary input.
  - **BaseFrame:** Using the text of the first frame as secondary input.
  - **VideoText:** Using the user's text input as secondary input throughout the sequence.
- **Falloff:** This hyperparameter controls the degree of variability by raising attention mappings to a power. Higher falloff values reduce areas of variation, leading to greater temporal stability but potentially limiting variance.

10.2. Grid Search Strategy

The grid search was primarily conducted on the SDXL model, utilizing a diverse set of prompts and systematically varying hyperparameters. Each configuration was evaluated using the following metrics:

- **Multi-Scale Structural Similarity (MS-SSIM):** Measures structural similarity between consecutive frames to evaluate content preservation.
- **Learned Perceptual Image Patch Similarity (LPIPS):** Analyzes perceptual similarity by comparing high-level features across frames.
- **Temporal Consistency Loss:** Assesses smoothness of motion using optical flow analysis.

For each configuration, these metrics were normalized to a [0, 1] range, and an equally weighted combined loss function was used for evaluation:

$$\begin{aligned}\text{Combined Loss} = {} & (1 - \text{Normalized MS-SSIM}) \\ & + \text{Normalized LPIPS} \\ & + \text{Normalized Temporal Consistency Loss}\end{aligned} \tag{8}$$

Lower combined loss values indicate better overall performance. We analyzed the most frequent high-performing hyperparameter configurations to identify optimal settings.

10.3. Results and Empirical Best Settings

From the grid search and manual tuning, the following configurations emerged as optimal for each model:

10.3.1 SDXL

- **Batch Size:** 3
- **Intersection Strategy:** Previous
- **Multi-Prompt Strategy:** VideoText
- **Guidance Scale:** 11.0
- **Falloff:** 2

10.3.2 SD1.5

- **Batch Size:** 3
- **Intersection Strategy:** Previous
- **Guidance Scale:** 11.0
- **Falloff:** 2

10.3.3 SD3

Manual testing revealed the following optimal settings for SD3:

- **Batch Size:** 2
- **Intersection Strategy:** Previous
- **Multi-Prompt Strategy:** VideoText / none
- **Guidance Scale:** 9.0 / 11.0
- **Falloff:** 1

10.4. Discussion

The consistency of effective hyperparameters across models highlights general principles for optimizing video generation in diffusion-based models:

- A **batch size of 3** achieves a balance between computational efficiency and temporal coherence.
- Using the **"Previous" intersection strategy** significantly enhances frame-to-frame continuity, reducing flickering and visual artifacts.
- A **guidance scale of 11.0** strikes an effective balance between adherence to text prompts and visual creativity.
- The **VideoText multi-prompt strategy** dynamically guides generation using the original text input and improves temporal consistency for supported architectures.
- **Falloff:** A falloff of 2 is ideal for SDXL and SD1.5, producing stable yet diverse outputs, whereas a falloff of 1 is better suited for SD3, maintaining sufficient variability.

These findings provide a robust framework for optimizing diffusion models for video generation tasks and offer a foundation for further experimentation and refinement.

11. IP-Adapter

In this section, we discuss the rationale for testing the IP-Adapter within our framework and evaluate its impact on video generation quality.

11.1. Rationale for Using IP-Adapter

The IP-Adapter [47] was integrated into our pipeline to leverage its cross-attention mechanism, which aligns with our modular and conditional generation objectives. As a well-established method in conditional image generation, the IP-Adapter provides fine-grained control over generated content by incorporating auxiliary inputs through attention mechanisms. This modular approach is more accessible than the architectural changes made by previous works in this area.

11.2. Results with IP-Adapter

The performance impact of the IP-Adapter is summarized in Tab. 1. Key observations include:

- **LPIPS:** A slight improvement was observed, with scores improving from 0.33 ± 0.1 (without IP-Adapter) to 0.316 ± 0.089 (with IP-Adapter). This suggests a marginal enhancement in perceptual quality.
- **MS-SSIM:** A modest increase in structural similarity was noted, with scores rising from 0.63 ± 0.137 (without IP-Adapter) to 0.655 ± 0.13 (with IP-Adapter).
- **Temporal Consistency Loss:** Negligible changes were observed, indicating that the IP-Adapter had limited impact on improving frame-to-frame coherence.

While these results highlight minor improvements in perceptual quality and structural similarity, the observed gains fall within the standard deviation, raising questions about their statistical significance.

11.3. Discussion on Results

Although the IP-Adapter provided minor enhancements in certain metrics, the improvements were not substantial enough to justify the added complexity it introduces into the pipeline. Given the lack of significant impact on temporal consistency and the marginal nature of the improvements, we conclude that the IP-Adapter may not be well-suited for our specific zero-shot video generation framework.

12. CLIP Results

Table 4. Quantitative comparison of CLIP Score for our method and previous works.

Method	Pre-Trained Model	CLIP Score
DirecT2V	**SD1.5**	0.276 ± 0.025
Free-Bloom	**SD1.5**	0.271 ± 0.022
T2V-Zero	**SD1.5**	0.294 ± 0.026
Ours	**SD1.5**	0.278 ± 0.03
Ours	**SD1.5 w/IP-Adapter[47]**	0.271 ± 0.034
Ours	**SD3**	0.276 ± 0.028
Ours	**SDXL**	0.271 ± 0.031

In this section, we present the CLIP scores for all models used in our main qualitative experiments. The results, summarized in Tab. 4, reveal minimal variance in CLIP scores across different models and configurations. While CLIP scores effectively measure text-image alignment, they do not correlate strongly with video generation performance or quality.

12.1. Analysis of CLIP Scores

As shown in Tab. 4, the CLIP scores for all models and configurations have vary little variance between them. Key observations include:

- **SD1.5-based models:** Scores ranged from 0.271 ± 0.022 (Free-Bloom) to 0.294 ± 0.026 (T2V-Zero). Our proposed method achieved scores of 0.278 ± 0.03 and 0.271 ± 0.034 across different configurations.
- **Newer models:** Both SD3 and SDXL achieved comparable scores, with 0.276 ± 0.028 and 0.271 ± 0.031, respectively.

Noting that the video output of these models was significantly different, these results demonstrate that while CLIP scores effectively fail to capture essential aspects of video quality, such as temporal coherence and perceptual fidelity. To address this we propose the three metrics we use. Details can be found in the main text.

13. User Study Setup

This section details the setup and execution of the user study conducted to validate the comparative performance of our video generation models.

13.1. Study Design

The user study was designed to evaluate the performance of our SD1.5 model against other SD1.5-based baseline models using 50 video prompts. Participants were asked to assess the generated videos across four evaluation criteria:

1. **Smoothness (Temporal Coherence):** The quality of transitions between frames, avoiding jumps or awkward motion.
2. **Picture Quality (Fidelity):** The visual fidelity and clarity of the video frames.
3. **Adherence to Description (Semantic Coherence):** How accurately the video aligned with the given text prompt.
4. **Overall Quality:** A holistic evaluation incorporating all three criteria.

Each video prompt was presented as four GIFs, corresponding to outputs from different models. The GIFs were randomly assigned labels (A, B, C, D) to eliminate potential biases. Participants ranked the GIFs for each evaluation criterion in descending order of preference (e.g., if A is preferred ABCD).

13.2. Study Implementation

The study was implemented as an interactive web application, allowing participants to evaluate videos in a structured and intuitive manner. The code for this web app will also be made public with the rest of the code. Key features of the study setup included:

- **Randomized Presentation:** GIFs for each video prompt were shuffled and assigned randomized labels for each participant.
- **Ranking Interface:** A simple ranking system required participants to assign a unique rank (1 to 4) to each GIF for all four criteria.
- **Data Collection:** Responses were validated to ensure completeness (e.g., each letter A, B, C, and D appeared exactly once per ranking) and stored in CSV format for aggregation and analysis.

Clear instructions were provided to ensure participants understood the evaluation process and the significance of each criterion.

13.3. Participant Details

A total of eight participants were involved in the study. Each participant evaluated all 50 video prompts across the four criteria, resulting in a total of 1,600 individual rankings. Participants represented a mix of technical and non-technical backgrounds, from ages 17 to 55, ensuring a balanced perspective on video quality.

13.4. Analysis and Observations

The rankings were aggregated across participants to derive average scores and identify trends. Key observations included:

- **High Variability in Preferences:** Standard deviations across rankings were consistently around 1 for all evaluation criteria, highlighting subjective variability in participant preferences.

- **Aggregated Insights:** Despite individual differences, the aggregated results consistently favored our model in terms of smoothness, picture quality, and adherence to descriptions.

Given the observed variability, we focused on aggregated rankings and qualitative trends rather than standard deviation as a primary metric.

13.5. Conclusion

The user study highlighted the strengths of our SD1.5 model in generating videos with superior smoothness, picture quality, and adherence to prompts compared to baseline models. While the small participant pool and the subjective nature of rankings introduced variability, the overall trends were consistent. Future studies involving a larger and more diverse participant base could further validate and refine these findings.

14. Additional Technical Details

We used an 8B LLaMA [7] model locally for prompt generation due to its practicality, but we also tested Qwen 2.5 7B [45] and Mistral 7B [22] (see Tab. 5). As our model is designed to be LLM-agonistic, there were no significant differences in performance. Naturally, the in-context information may need to be optimized for each model, but in general, the LLaMa model performed best, which is why we used it in our main testing.

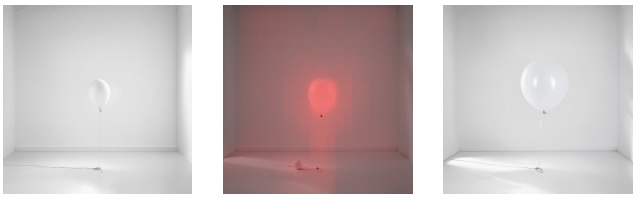
Our model also does not depend on any particular ODE solver; as such, we used the standard options provided in the Diffusers Library [43].

We do not fix the seeds across models, as their internal sampling mechanisms can yield differing outputs even with a fixed seed. Fig. 6 demonstrates that distinct methods can produce substantially different results despite fixed seeds (and the same image generator).

Fig. 5 provides a detailed example of how the attention mechanism works. It shows an example of the different text components and how they are combined with a CLIP model to generate an attention map over the previous frame. This attention map highlights areas that require high variance. This allows the image generator to make more changes in the given region, and as we can see, the balloon has changed in the next frame.

Table 5. Quantitative performance of EIDT-V using alternative LLMs. For more details, please refer Tab. 1.

Method	Pre-Trained Model	Unmodified Architecture	MS-SSIM (↑)	LPIPS (↓)	Temporal Consistency (↓)
EIDT-V	**SD1.5 w/ Qwen LLM**	✓	0.572 ± 0.151	0.370 ± 0.093	0.168 ± 0.058
EIDT-V	**SD1.5 w/ Mistral LLM**	✓	0.599 ± 0.122	0.353 ± 0.077	0.162 ± 0.061



Before Attention Map Next Frame

Fixed part: *Scene: Empty white background; a balloon tied to a string. Character: None. Setting: Indoor location; dim lighting; no distractions. Background: Simple, plain surface; minimal reflections.*
Previous Dynamic part: *Inflation begins slow.*
New Dynamic part: *Volume increases gradually.*
Differences Detected: ["Balloon inflating", "Volume increasing"]

Figure 5. Our method detects differences, generates attention map and combines them by taking the maximal value at each pixel. Bright red regions in attention correspond to high variance.

15. Large Scale Changes

Extreme scene changes (e.g., when the subject moves forward while the background moves in the opposite direction) are challenging for all training-free approaches. As shown in Fig. 6, methods such as T2VZero and DirecT2V often fail to preserve the subject adequately, while FreeBloom exhibits excessive variation. In contrast, our method localizes changes, effectively balancing consistency and variance.

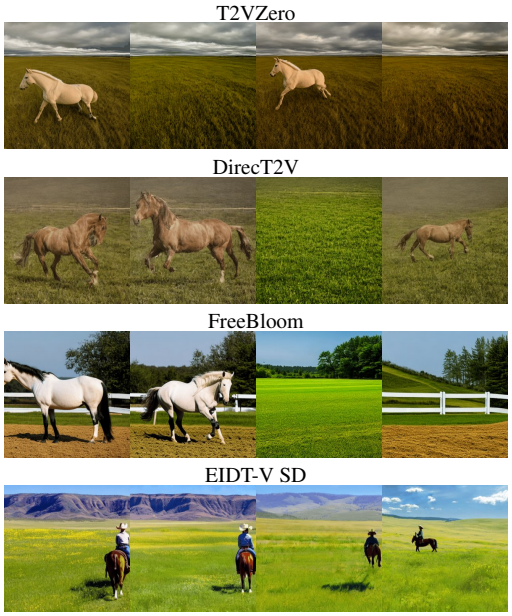


Figure 6. Qualitative comparison SD1.5 based video-generation models for the prompt: "**A first-person view from atop a horse, its ears and mane visible, moving forward across a grassy field**". A fixed seed was used across all models.

16. Additional Qualitative

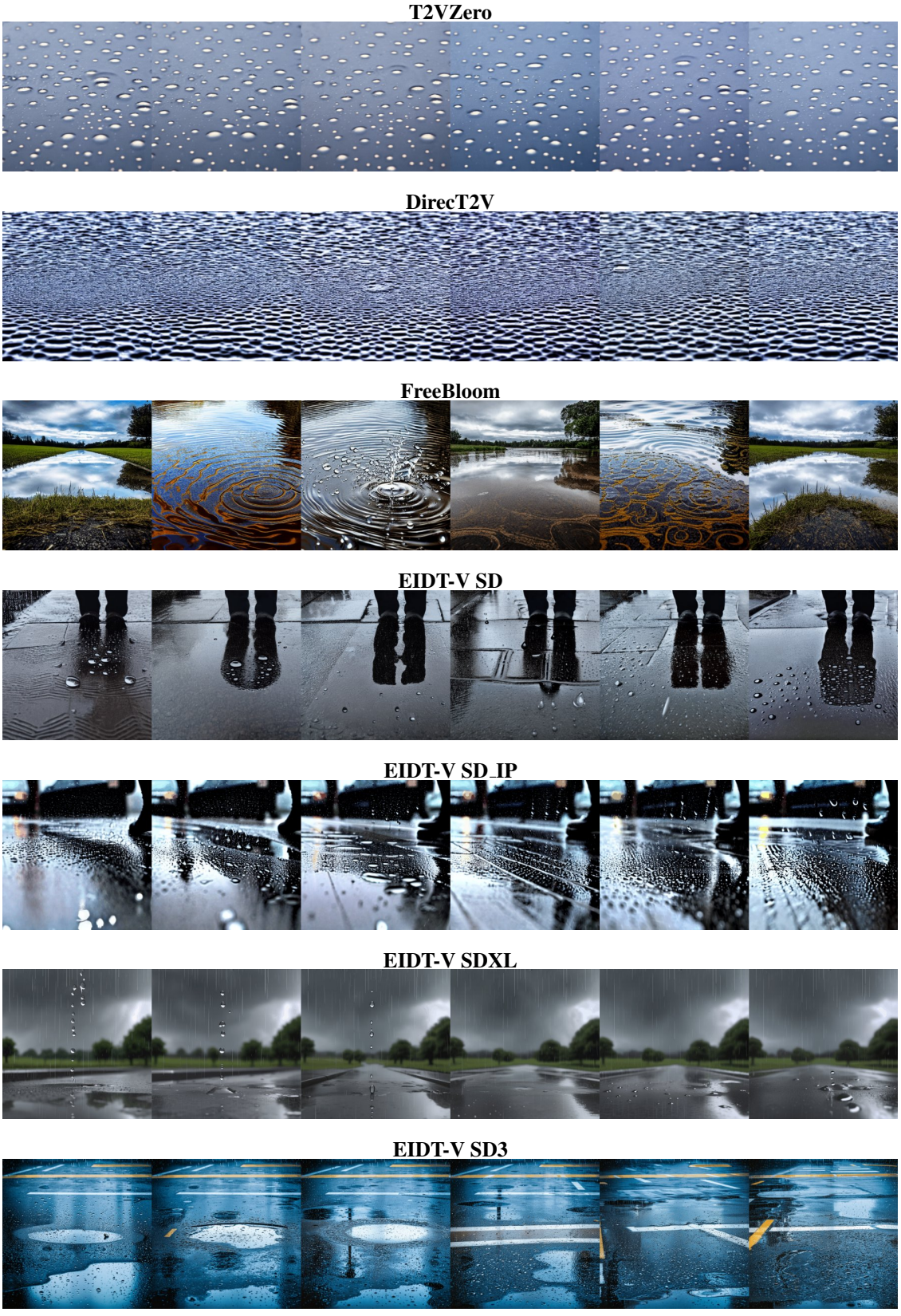


Figure 7. Raindrops falling into a puddle creating ripples.

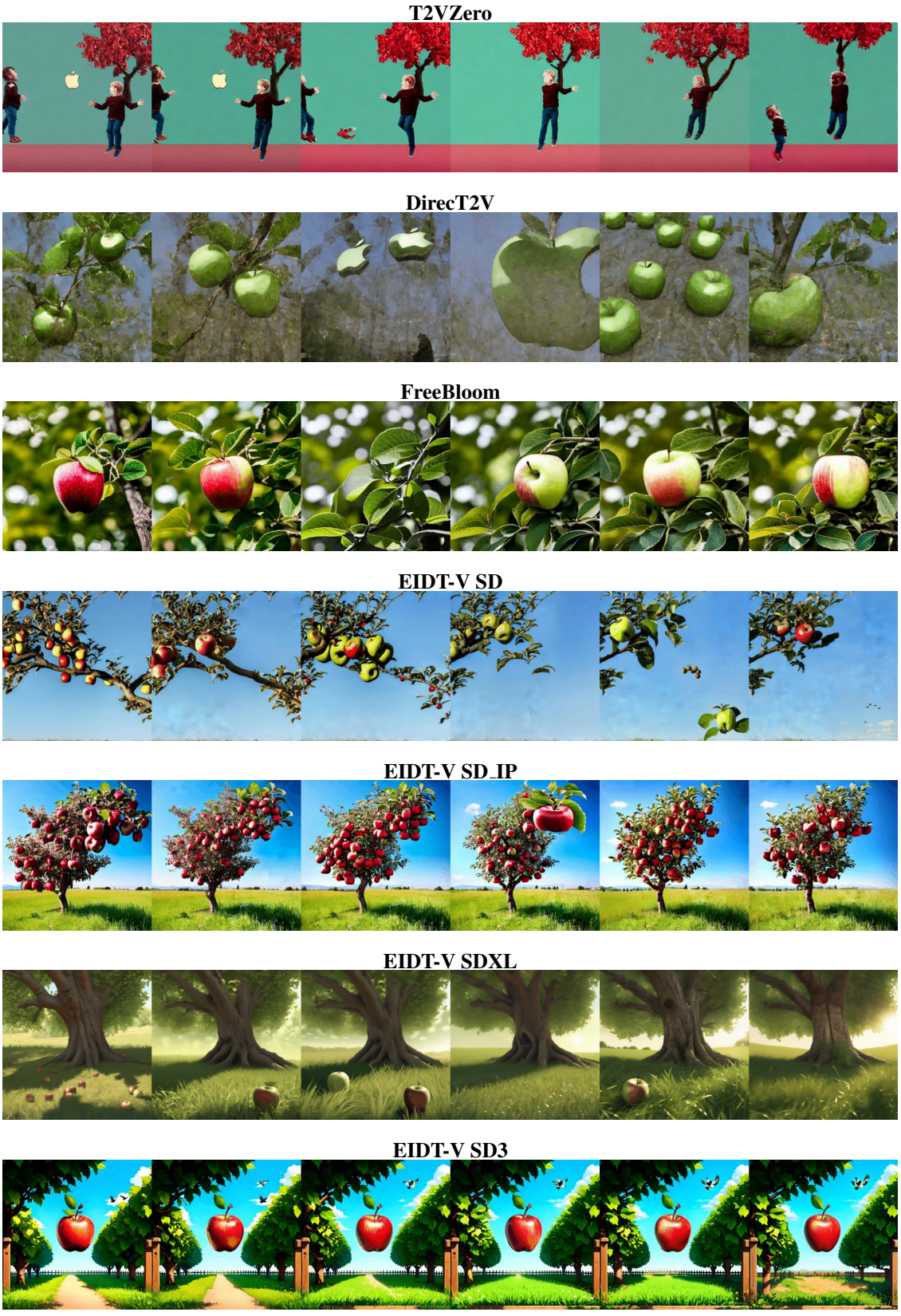


Figure 8. An apple falling from a tree and bouncing on the ground.

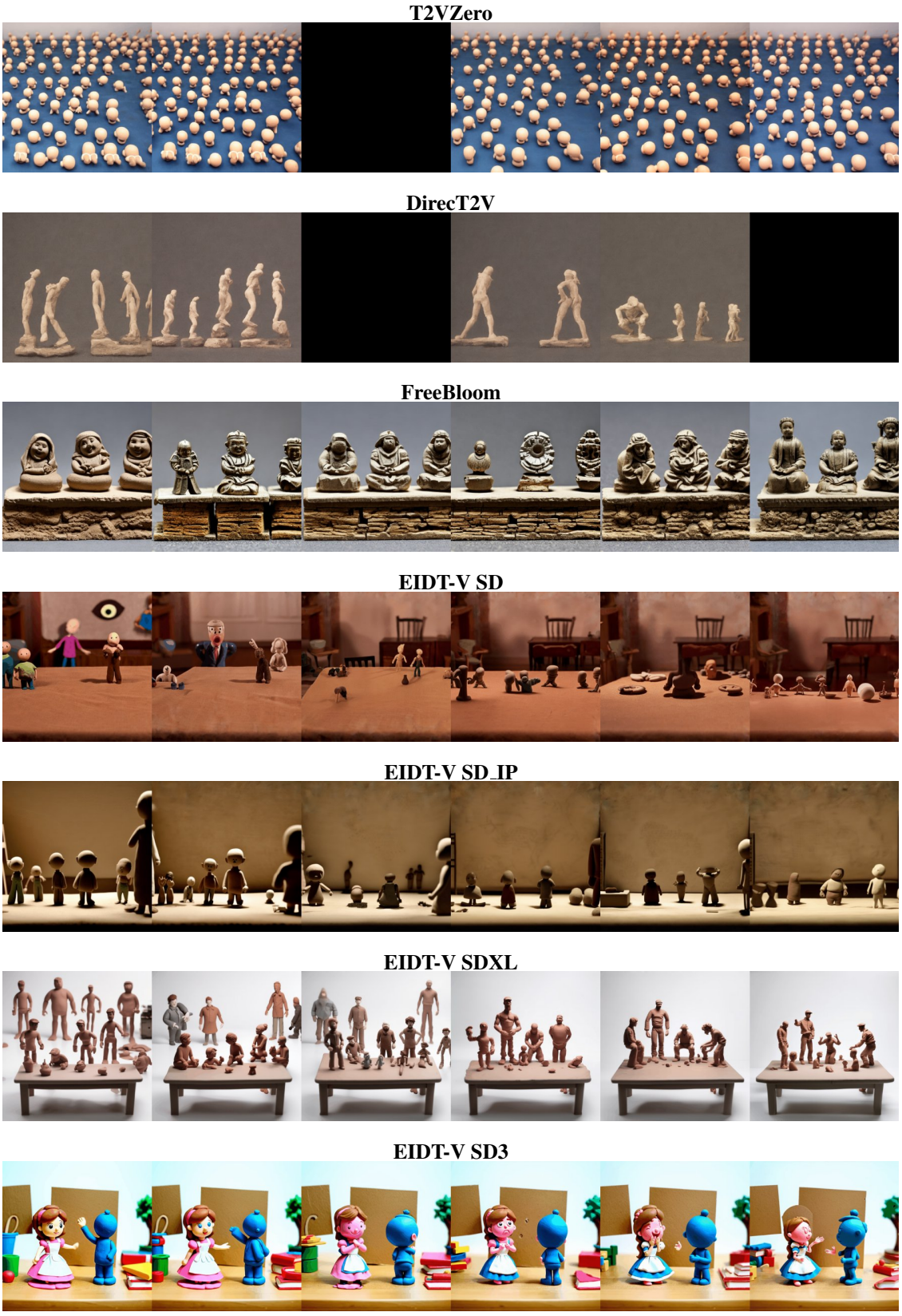


Figure 9. A stop-motion animation of clay figures moving.

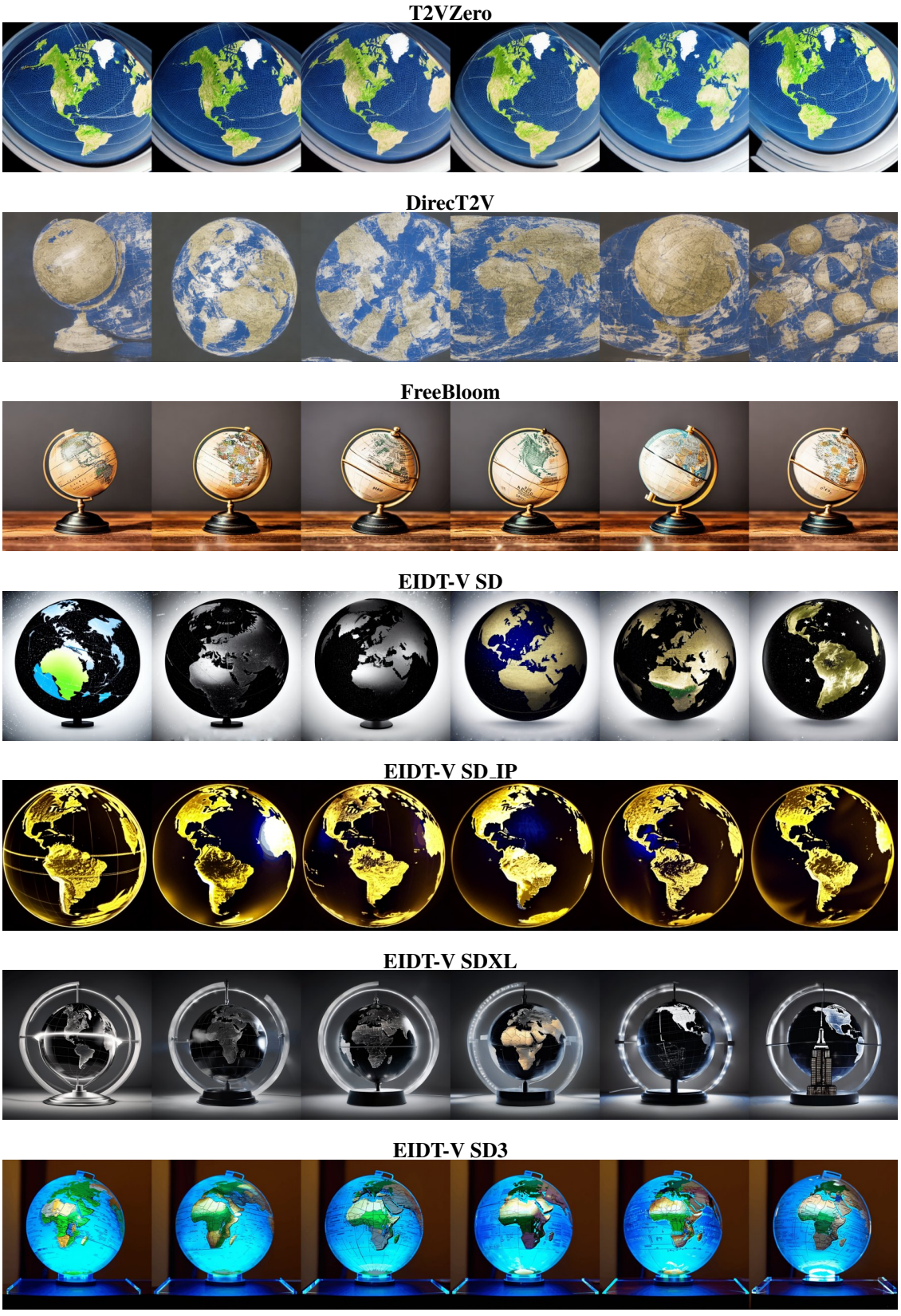
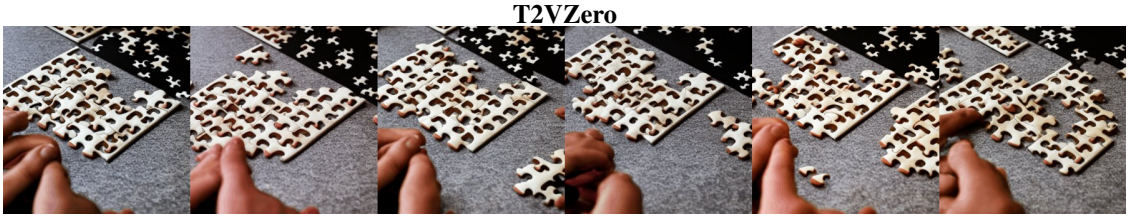


Figure 10. A spinning globe showing continents passing by.

T2VZero

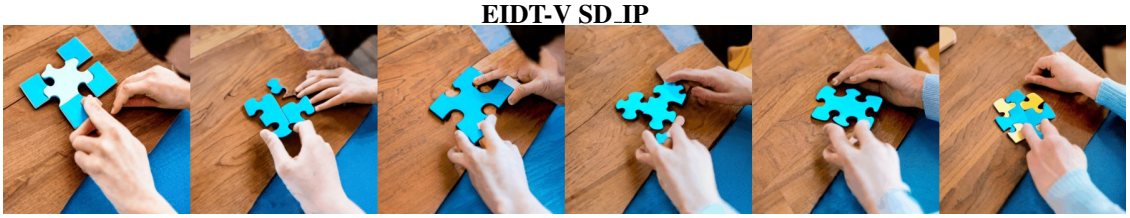


DirecT2V

FreeBloom

EIDT-V SD

EIDT-V SD_IP



EIDT-V SDXL

EIDT-V SD3

Figure 11. A puzzle being assembled piece by piece.

17. Additional Best

Here we highlight some of our best generations using the more powerful models SDXL and SD3.

17.1. SDXL

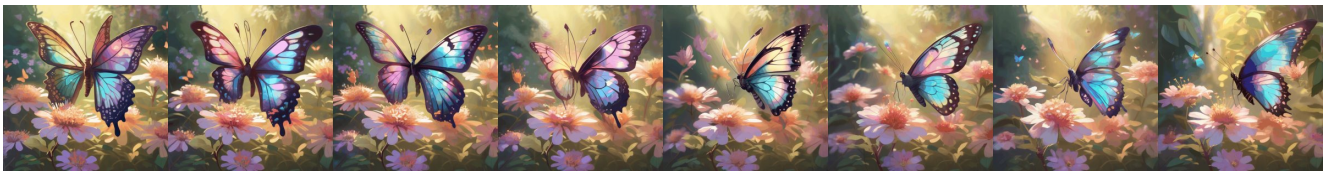


Figure 12. A butterfly gently flapping its wings while resting on a flower.

Figure 13. A figure skater gliding across an ice rink with smooth turns.

Figure 14. A galaxy swirling with stars and nebulae in deep space.

Figure 15. A lightning bug flying through a dark meadow.

Figure 16. A musician playing a slow, peaceful tune on an acoustic guitar.

Figure 17. A phoenix slowly rising from glowing embers.

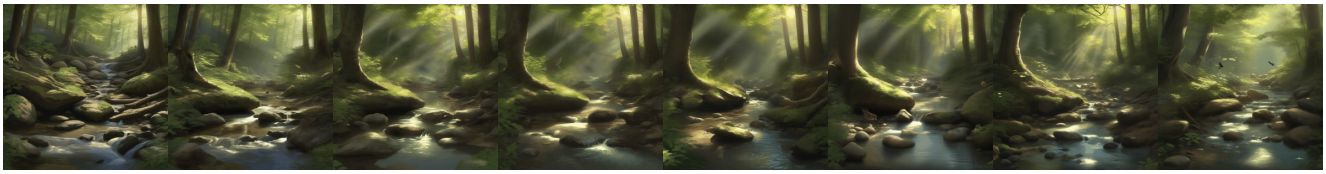


Figure 18. A stream flowing slowly over rocks in a forest.

17.2. SD3 Medium

Figure 19. A butterfly gently flapping its wings while resting on a flower.

Figure 20. A dolphin gracefully gliding through turquoise waves.

Figure 21. A dragon breathing a gentle stream of smoke from its nostrils.

Figure 22. A family of penguins huddling together in a snowstorm.

Figure 23. A musician playing a slow, peaceful tune on an acoustic guitar.

Figure 24. A person writing slowly in a journal with an ink pen.

Figure 25. A portal opening and closing slowly in a mystical cave.

Figure 26. A squirrel nibbling on an acorn under a tree.

Figure 27. A unicorn grazing in a meadow under a rainbow.

Figure 28. Golden leaves swirling softly in the autumn wind.