A. Implementation Details for T2V and STIV

Given that we use spatial-temporal attention, we first pre-train the T2I model using only an image dataset. Subsequently, we load the EMA weights from the T2I model, excluding the temporal attention. In our work, we use the per-frame VAE, which is the same one used in the T2I model. On top of that, we use a temporal patch of size 2 in the DiT part for video models. We modify the T2I cubify weights by inflating the 3D convolution weight in the temporal dimension. For video training data, we select one frame from every three frames and add independent and identically distributed Gaussian noise to each frame. Following standard practice, we randomly replace text prompt with empty string 10% during training. In our STIV setting, we also independently randomly drop image condition 8% during training. For both T2V and STIV models, the CFG scale is set to 7.5. The training schedule follows the progressive training recipe described in section 2.3.

B. Implementation Details of Text Encoders

We used our internal CLIP text encoder to encode text into embeddings. Concretely, a text is first tokenized via a T5 tokenizer. The tokenized text is mapped into embeddings via an embedding lookup table and further encoded via 32 layers of transformer with casual attention. Each transformer layer contains 20 attention heads. Each attention head has 64 hidden dimensions. The output text embedding has a dimension of 1280.

C. Ablation Study on T2I Generation

Baseline Setup For our base model, we employed the PixArt-α architecture [8], which builds on the DiT [43] model with added cross-attention layers to integrate image tokens with text embeddings. As pre-trained components, we used the open-source sd-vae-ft-ema model[4] and OpenAI CLIP L14 model[5], both of which are widely adopted in the community. We conduct our experiments using the XL model configuration with a 256^2 image size. The full baseline model, which includes the VAE and CLIP text encoder, has approximately 1.06 billion parameters. For noise generation and denoising, we used a diffusion-based approach with Stable Diffusion's default noise schedule. The training was conducted with a batch size of 4,096 over 400k steps, which corresponds to approximately 1.4 epochs on our internal text-to-image dataset.

Table 5 summarizes the results of our ablation study, focusing on the following aspects:

Stabilized Training Leveraging recent advancements in LLM and diffusion model architectures, we integrated QK-Norm [27] to manage the activation scale within attention layers. Additionally, we applied sandwich-norm [24] to both the inputs and outputs of the attention layer and the feedforward layer. Projected conditions, including timestep embeddings, pooled CLIP text embeddings, and micro condition embeddings, were normalized before being input to AdaLN. These normalization techniques enhanced training stability, allowing us to increase the learning rate from 1×10^{-4} to 2×10^{-4}, and also improved quality.

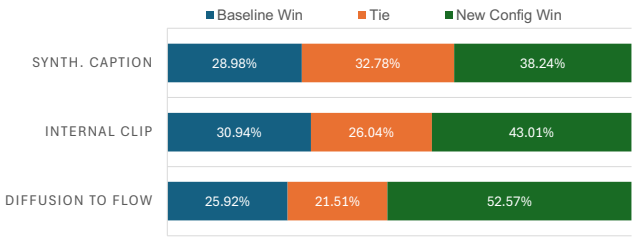


Figure 4. Human evaluation results on significant changes in T2I ablation study Tab. 5.

Noising and Denoising Process Formulation We explored optimized noising/denoising formulations by replacing the diffusion process with a flow-based linear interpolant[41]. Additionally, we applied renormalization at each inference step to counteract potential side effects from high classifier-free guidance (CFG) values. Here, the norm of the prediction with CFG was linearly scaled to match the conditional prediction norm, as explained in Sec. 2.2.

Training Cost Optimization To reduce training costs, we evaluated three strategies: (1) switching from the AdamW optimizer to Adafactor, (2) applying MaskDiT training with a 50% masking ratio, and (3) using a shared AdaLN module across layers instead of unique instances per layer. These changes reduced per-device HBM usage from approximately 28GB to 11GB, allowing us to train on v5e TPUs instead of the more costly v5p TPUs. Notably, as shown in Table 5, masked training may adversely affect metrics such as FID and HPS. However, we found additional unmask finetuning for a short duration (e.g. 50k steps) can fix the artifacts causing these score drops. However, this additional training phase was not included in the final configuration, as further training on video generation can address this issue as well.

Enhanced Pre-trained Models and Conditioning We evaluated improvements from advanced pre-trained models and additional conditioning techniques. Specifically, we upgraded from the OpenAI CLIP L14 to an internally trained CLIP-bigG model [34] and from a 4-channel to an 8-channel VAE. We also introduced 2D RoPE to support masked training and added micro-conditions, inspired by SDXL [45], to mitigate cropping artifacts in elongated ob-

[4]https://huggingface.co/stabilityai/sd-vae-ft-ema
[5]https://huggingface.co/openai/clip-vit-large-patch14

jects. Finally, including synthetic captions generated via [33] in training data results in notable performance gains.

Human Evaluation of Model Changes To validate improvements observed in automated metrics, we conducted human evaluations for key modifications, including the addition of synthetic captions, upgrade of CLIP model, and transition from diffusion to flow matching based objective. Human raters are asked to asses image fidelity, text-image alignment, and visual appeal, and give 5 level preference ratings for image pairs. Each pair is sent to 5 raters for rating and the image pair will be considered tie of combined voting is neutral. Results from Figure 4 demonstrate clear alignment between automated metrics and human judgments. This justifies to use automatic evaluation as development metrics to maintain generation quality and prevent regressions leading to significant quality losses.

C.1. Video Data Engine

Data quality is pivotal for video generation models. However, curating large-scale, high-quality datasets remains challenging due to issues like noisy captions, hallucinations, and limited diversity in video content and duration. To address these concerns, we propose a **Video Data Engine** (Fig. 5)—a comprehensive pipeline that improves dataset quality and reduces hallucinations, ultimately enhancing model performance. More details can be found in Sec. **??** in the appendix.

Our approach focuses on three key questions: (1) How to preprocess raw videos for better consistency? (2) What is the effect of data filtering on model performance? (3) How can advanced video captioning reduce hallucinations and improve outcomes? We use Panda-70M [9] as a working example and produce a curated subset, Panda-30M, via our pipeline.

Video Pre-processing and Feature Extraction. We employ PySceneDetect[6] to remove abrupt transitions and inconsistent segments, yielding more coherent clips. We then extract key features (e.g., motion and aesthetic scores) to guide subsequent filtering.

Data Engine for Filtering Effective data filtering is crucial for improving dataset quality and reducing hallucinations. We develop an automated filtering infrastructure that supports efficient data selection, quality control, and continuous improvement throughout the model's development lifecycle. For instance, we can sample high-quality videos with predefined resolutions / motion scores for the fine-tuning stage. This filtering system allows us to systematically remove low-quality videos and focus on data that enhances model performance. From Panda-30M, we further apply filtering based on motion score and aesthetic score to obtain Panda-10M, named as a high-quality version of Panda-30M. The results are summarized in Tab. 12: instead of pursuing data volume, higher-quality videos have the potential to achieve more promising results.

[6] https://github.com/Breakthrough/PySceneDetect

Video Captioning Model High-quality video-text pairs are essential for training text-to-video models. Existing datasets often suffer from noisy or irrelevant captions, limited in describing temporal dynamics. We initially attempted a frame-based captioning approach followed by LLM summarization [3], but found that single-frame captions fail to represent motion, and LLM summarization can induce hallucinations. To improve caption quality while balancing cost, we employ LLaVA-Hound-7B [67]—a video LLM capable of producing more coherent and motion-aware descriptions.

Caption Evaluation and Ablations To objectively assess caption accuracy, we introduce **DSG-Video**(Fig. 6), a module inspired by DSG [11], that detects hallucinated objects by probing captions with LLM-generated questions and verifying object presence in sampled video frames using a multimodal LLM. This yields two metrics, DSG-Video$_i$ and DSG-Video$_s$, reflecting hallucination at the object and sentence levels, respectively. We compare two captioning strategies—frame-based plus LLM summarization (FCapLLM) and direct video captioning (VCap)—on the Panda-30M dataset. As shown in Tab. 11, VCap reduces hallucinations and increases the diversity of described objects, leading to improved T2V model performance. These results show that richer, more accurate video descriptions can significantly enhance downstream generation quality.

D. Detailed Results for T2V and STIV

D.1. VBench and VBench-I2V Evaluation Metrics

We follow the same as the evaluation protocol provided by VBench [30].

D.1.1. Video Quality

Video Quality is divided into two aspects: *Temporal Quality* and *Image Quality*. Temporal Quality evaluates cross-frame consistency, including (1) *Subject Consistency*, ensuring that subjects maintain a consistent appearance across frames; (2) *Background Consistency*, assessing stability in the background using feature similarity; (3) *Temporal Flickering*, measuring smooth transitions in both static and dynamic areas; (4) *Motion Smoothness*, evaluating the fluidity and realism of motion; and (5) *Dynamic Degree*, analyzing the presence of large-scale dynamics or motions. Image Quality focuses on individual images and evaluates (1) *Aesthetic Quality*, considering artistic appeal and visual richness, and (2) *Imaging Quality*, measuring clarity, noise, and other distortions.

Caption	Total Object	DSG-Video$_i$ (↓)	DSG-Video$_s$ (↓)	MSRVTT FVD (↓)	VBench (↑)
FCapLLM	1249	6.4	24.0	808.1	64.2
VCap	1911	**5.3**	**15.0**	**770.9**	**65.6**

Table 11. Compare different captions using XL T2V model. DSG-Video metrics are calculated from 100 random captions.

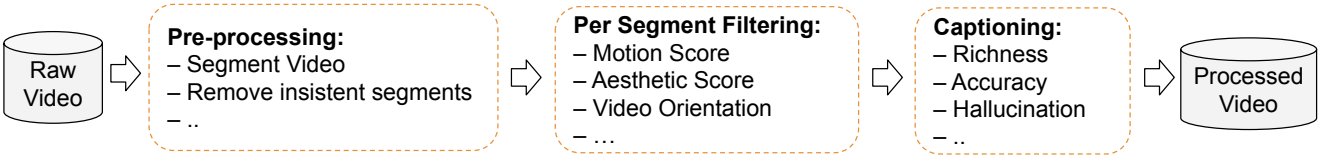


Figure 5. An overview of our video data engine, including video pre-processing, filtering, and video captioning.

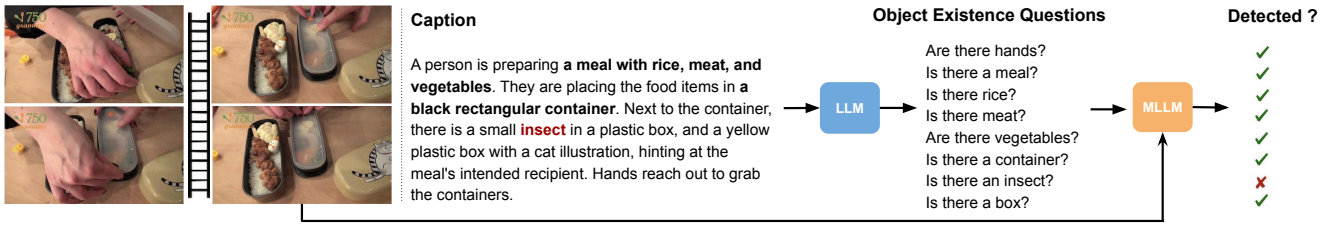


Figure 6. An overview of DSG-Video's approach to detecting object hallucinations in captions: we use an LLM to generate questions and another MLLM to validate the presence of the object across frames. If the MLLM fails to detect the object in all frames, the object is classified as a hallucination.

Data	MSRVTT	VBench		
	FVD ↓	Quality ↑	Semantic ↑	Total ↑
Panda-30M	770.9	80.4	**73.6**	65.6
Panda-10M	**759.2**	**80.8**	73.4	**66.2**

Table 12. Compare Panda-30M and Panda-10M (high-quality) using XL T2V model.

D.1.2. Video-Condition Consistency

Video-Condition Consistency ensures alignment with the input prompt and is categorized into *Semantics* and *Style*, each with finer-grained dimensions.

Semantics (1) *Object Class*: Measures the success of generating specific objects described in the text prompt. (2) *Multiple Objects*: Evaluates the ability to compose multiple objects from different classes in a single frame. (3) *Human Action*: Assesses whether the generated video accurately captures actions described in the prompt. (4) *Color*: Ensures synthesized object colors align with the text description. (5) *Spatial Relationship*: Checks whether spatial relationships between objects align with the prompt. (6) *Scene*: Evaluates consistency between generated scenes and the intended description (e.g., "ocean" versus "river").

Style (1) *Appearance Style*: Measures consistency of styles mentioned in the prompt, such as "oil painting" or "cyberpunk" (2) *Temporal Style*: Assesses temporal continuity of styles across frames, ensuring smooth transitions.

Overall Consistency We further evaluate *Overall Consistency* using metrics that combine semantic and style alignment, reflecting both the accuracy and coherence of generated videos.

VBench-I2V builds upon the VBench with three new Video-Image Alignment metrics: Subject Consistency, Background Consistency, and Camera Motion Control. These additional metrics provide a more comprehensive evaluation by focusing on how well the generated video aligns with the input image and specified prompt instructions. Specifically, Subject Consistency evaluates the alignment between the subject in the input image and the generated video, ensuring coherence in character or object representation. Background Consistency assesses the continuity of the background scene between the input image and the video, highlighting the model's ability to maintain a consistent environment. Camera Motion Control, under Video-Text Alignment, examines the adherence to camera control directions as described in the prompt, which is crucial for generating realistic video sequences that respond to specified dynamic instructions.

D.2. Detailed Results on VBench and VBench-I2V

We showcase the detailed version of the performance shown in Tab. 13 and Tab. 14.

Model	Subject Cons.	Back. Cons.	Temporal Flickering	Motion Smooth.	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class	Multiple Objects	Human Action
CogVideoX-5B [63]	96.2	96.5	98.7	96.9	80.0	62.0	62.9	85.2	62.1	99.4
CogVideoX-2B [63]	96.8	96.6	98.9	97.7	59.9	60.8	61.7	83.4	62.6	98.0
Allegro [71]	96.3	96.7	99.0	98.8	55.0	63.7	63.6	87.5	59.9	91.4
AnimateDiff-V2 [25]	95.3	97.7	98.8	97.8	40.8	67.2	70.1	90.9	36.9	92.6
OpenSora V1.2 [70]	96.8	97.6	**99.5**	98.5	42.4	56.9	63.3	82.2	51.8	91.2
T2V-Turbo [36]	96.3	97.0	97.5	97.3	49.2	63.0	**72.5**	94.0	54.7	95.2
VideoCrafter-2.0 [7]	96.9	98.2	98.4	97.7	42.5	63.1	67.2	92.6	40.7	95.0
LaVie-2 [59]	97.9	98.5	98.8	98.4	31.1	**67.6**	70.4	**97.5**	64.9	96.4
LaVIE [59]	91.4	97.5	98.3	96.4	49.7	54.9	61.9	91.8	33.3	96.8
ModelScope [58]	89.9	95.3	98.3	95.8	66.4	52.1	58.6	82.2	39.0	92.4
VideoCrafter [6]	86.2	92.9	97.6	91.8	**89.7**	44.4	57.2	87.3	25.9	93.0
CogVideo [29]	92.2	95.4	97.6	96.5	42.2	38.2	41.0	73.4	18.1	78.2
PIKA [44]	96.9	97.4	**99.7**	**99.5**	47.5	62.4	61.9	88.7	43.1	86.2
Gen-3 [52]	97.1	96.6	98.6	99.2	60.1	63.3	66.8	87.8	53.6	96.4
Gen-2 [51]	97.6	97.6	99.6	**99.6**	18.9	67.0	67.4	90.9	55.5	89.2
KLING [32]	**98.3**	97.6	99.3	99.4	46.9	61.2	65.6	87.2	68.1	93.4
EMU3 [23]	95.3	97.7	98.6	98.9	79.3	59.6	62.6	86.2	44.6	77.7
XL	96.0	98.5	98.4	96.5	62.5	56.3	59.3	91.5	41.3	98.0
XXL	97.5	**98.9**	99.1	98.2	48.6	56.2	59.7	91.1	49.1	**99.0**
M-256	96.0	98.5	98.6	97.2	68.1	57.0	60.8	88.8	62.1	98.0
M-512	95.9	96.9	98.8	98.0	59.7	60.6	62.5	85.9	**72.4**	96.0
M-512-SFT	96.7	97.4	98.7	98.3	70.8	61.7	63.9	88.1	67.7	97.0
M-512-SFT+TUP	94.8	95.9	98.7	99.2	70.8	63.7	65.0	88.9	70.3	95.0
M-512-UnMSFT	94.3	96.9	98.8	96.7	77.8	61.4	68.6	90.0	72.3	97.0
M-512-UnMSFT+TUP	95.2	95.8	98.8	99.2	70.8	63.6	65.9	90.0	69.8	94.0

Model	Color	Spatial Rel.	Scene	App. Style	Temp. Style	Overall Cons.	Quality Score	Semantic Score	Total Score	Averaged Scores
CogVideoX-5B [63]	82.8	66.4	53.2	24.9	25.4	27.6	82.8	77.0	81.6	70.0
CogVideoX-2B [63]	79.4	69.9	51.1	24.8	24.4	26.7	82.2	75.8	80.9	68.3
Allegro [71]	82.8	67.2	46.7	20.5	24.4	26.4	83.1	73.0	81.1	67.5
AnimateDiff-V2 [25]	87.5	34.6	50.2	22.4	26.0	27.0	82.9	69.8	80.3	64.7
OpenSora V1.2 [70]	90.1	68.6	42.4	24.0	24.5	26.9	81.4	73.4	79.8	66.0
T2V-Turbo [36]	89.9	38.7	55.6	24.4	25.5	28.2	82.6	74.8	81.0	67.4
VideoCrafter-2.0 [7]	92.9	35.9	55.3	**25.1**	25.8	28.2	82.2	73.4	80.4	66.0
LaVie-2 [59]	91.7	38.7	49.6	**25.1**	25.2	27.4	83.2	75.8	81.8	67.6
LaVIE [59]	86.4	34.1	52.7	23.6	25.9	26.4	78.8	70.3	77.1	63.8
ModelScope [58]	81.7	33.7	39.3	23.4	25.4	25.7	78.1	66.5	75.8	62.4
VideoCrafter [6]	78.8	36.7	43.4	21.6	25.4	25.2	81.6	72.2	79.7	62.3
CogVideo [29]	79.6	18.2	28.2	22.0	7.8	7.7	72.1	46.8	67.0	52.3
PIKA [44]	90.6	61.0	49.8	22.3	24.2	25.9	82.9	71.8	80.7	66.1
Gen-3 [52]	80.9	65.1	54.6	24.3	24.7	26.7	84.1	75.2	82.3	68.5
Gen-2 [51]	89.5	66.9	48.9	19.3	24.1	26.2	82.5	73.0	80.6	66.1
KLING [32]	89.9	**73.0**	50.9	19.6	24.2	26.4	83.4	75.7	81.9	68.8
EMU3 [23]	88.3	68.7	37.1	20.9	23.3	24.8	84.1	68.4	81.0	66.7
XL	86.4	42.4	54.4	22.4	26.3	27.8	80.7	72.5	79.1	66.1
XXL	90.8	45.1	45.5	22.1	26.1	27.4	81.2	72.7	79.5	65.9
M-256	83.6	44.5	54.7	22.5	26.6	28.4	82.7	74.8	80.6	67.9
M-512	91.2	51.0	53.6	23.9	25.8	27.8	82.2	77.0	81.2	68.8
M-512-SFT	93.7	58.0	52.8	24.6	26.2	28.5	83.9	78.3	82.8	70.3
M-512-SFT+TUP	**94.7**	50.6	**57.3**	24.5	**26.7**	28.6	84.2	78.5	**83.1**	70.3
M-512-UnMSFT	92.0	59.8	53.1	24.8	**26.7**	**28.8**	83.7	**79.5**	82.9	**71.2**
M-512-UnMSFT+TUP	87.7	46.9	57.1	24.5	26.6	28.5	**84.4**	77.2	83.0	69.7

Table 13. Detailed Evaluation Results for Text-To-Video Generation Models.

E. Details of Model Initialization Ablations

To facilitate a fair comparison for different initialization methods we estimate the FLOPs associated with spatial-temporal computation in the transformer for various model training steps (Tables 15 and 16). When controlling for FLOPs we take into account, the compute used to pre-train the intermediate models, the reduction in an effective number of tokens due to masking in the relevant attention blocks, the increased parameter count when temporal attention is included, and the increased number of tokens passed to the model during high resolution training. For both the high resolution and higher frame count experiments we attebyto keep the compute budget across model initialization ablations similar. Tables 17 and 18 show the VBench quality metrics for high resolution and high frame count XL sized models respectively.

F. Study of T2V on Physics Commonsense Alignment Benchmark

We evaluated our models on physics commonsense benchmark VideoPhy, which outperforms both open sourced and close sourced models on the leaderboard, the results are shown in Tab. 19.

Model	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality
DynamicCrafter-256 [60]	94.7	98.3	98.1	97.8	40.6	58.7
DynamicCrafter-512 [60]	93.8	96.6	95.6	96.8	69.7	60.9
Animate-Anything [13]	98.9	98.2	98.1	98.6	2.7	67.1
SVD [3]	95.5	96.6	98.1	98.1	52.4	60.2
SEINE-512 [10]	95.3	97.1	97.3	97.1	27.1	64.6
VideoCrafter-I2V [7]	97.9	98.8	98.2	98.0	22.6	60.8
Consistent-I2V [49]	95.3	98.3	97.6	97.4	18.6	59.0
I2VGen-XL [68]	94.2	97.1	98.3	26.1	26.1	64.8
TI2V-M	95.4	98.9	97.2	98.1	32.1	59.0
TI2V-M-512	99.5	99.3	99.5	99.6	10.2	62.5
STIV-M-512	98.1	98.6	98.7	99.1	24.0	65.4

Model	Imaging Quality	I2V Subject	I2V Background	Camera Motion	I2V Quality	Final Score
DynamicCrafter-256 [60]	62.3	97.1	97.6	20.9	80.2	88.4
DynamicCrafter-512 [60]	68.6	97.2	97.4	32.0	81.6	89.1
Animate-Anything [13]	72.1	98.8	98.6	13.1	81.2	89.8
SVD [3]	69.8	98.8	98.6	62.3	82.8	89.9
SEINE-512 [10]	71.4	97.2	96.9	21.0	80.6	88.4
VideoCrafter-I2V [7]	71.7	91.2	91.3	33.6	81.3	85.1
Consistent-I2V [49]	66.9	95.8	96.0	33.9	78.9	86.8
I2VGen-XL [68]	69.1	96.5	96.8	18.5	81.2	88.5
TI2V-M	66.1	97.0	97.4	22.7	78.8	87.6
TI2V-M-512	71.5	99.2	97.3	13.2	82.1	90.1
STIV-M-512	71.0	98.8	97.5	15.1	81.9	89.8

Table 14. Detailed Evaluation Results for Text-Image-To-Video Generation Models.

Init. Method	Models	Stage 1	Stage 2	Stage 3	Stage 4	Total
Scratch	T2V-512	5.93				5.93
T2V-256	T2I-256, T2V-256, T2V-512	1.11	2.05	2.84		6.00
T2I-512	T2I-256, T2I-512, T2V-512	1.11	8.43	4.02		5.97
Both	T2I-256, T2V-256, T2I-512, T2V-512	1.11	2.05	8.43	1.98	5.98

Table 15. A breakdown of FLOPs for training high resolution T2V models. Unit 10^{21}.

Init. Method	Models	Stage 1	Stage 2	Stage 3	Total
T2I	T2I-256, T2V-256-40	1.11	2.05		3.16
T2V (int.)	T2I-256, T2V-256-20, T2V-256-40	1.11	1.02	1.02	3.16
T2V (ext.)	T2I-256, T2V-256-20, T2V-256-40	1.11	1.02	1.02	3.16
T2V 2x (int.)	T2I-256, T2V-256-20 2x stride, T2V-256-40	1.11	1.02	1.02	3.16

Table 16. A breakdown of FLOPs for training high frame count T2V models. Unit: 10^{21}.

G. Study of Class-to-Video on UCF-101

UCF-101 is an action recognition dataset, which contains 101 classes over 9.5K training videos. Here we train STIV from scratch and perform label-to-video (L2V) generation with 16 frames and 128^2 resolution. We follow TATS [21] to adopt the Inception Score (IS) [53] and FVD for the evaluation[7].

Tab. 20 shows that our L2V-XL achieves significant improvements, leading to +12% IS and -22% FVD over MAGVIT. This also highlights the effectiveness of our model design for convention video generation. From the ablation study over different modulations, only without spatial mask makes a lower FVD but degrades IS, while all other settings hurt the performance.

[7]Following our baselines (https://github.com/songweige/TATS/issues/13), we apply C3D [57] pre-trained on UCF-101 for the IS logits. For FVD, we adopt I3D [5] pre-trained on Kinetics-400 [31] to calculate the video embeddings.

Initial Method	Subject Cons.	Background Cons.	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
Scratch	**93.1**	97.1	97.9	97.3	**61.4**	58.6	58.6	87.0
T2V-256	91.9	97.1	98.0	**97.5**	58.6	59.4	59.7	**91.2**
T2I-512	92.3	97.2	98.2	97.0	52.2	60.0	59.3	88.8
Both	92.4	**97.3**	**98.3**	97.4	53.9	**60.7**	**60.6**	88.2
Initial Method	**Multiple Objects**	**Human Action**	**Color**	**Spatial Relationship**	**Scene**	**App. Style**	**Temp. Style**	**Overall Cons.**
Scratch	29.7	95.4	88.3	33.8	46.9	21.6	25.8	26.4
T2V-256	45.7	95.8	**89.0**	36.3	50.0	21.9	25.8	27.3
T2I-512	47.4	**96.4**	87.9	**37.0**	49.1	22.5	26.2	27.8
Both	**49.7**	96.0	88.1	36.7	**52.3**	**22.8**	**26.3**	**28.0**

Table 17. Detailed VBench metrics of different model initialization methods for higher resolution T2V model training.

Initial Method	Subject Cons.	Background Cons.	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
T2I	**93.2**	**98.1**	**98.7**	95.2	57.8	54.2	58.2	84.6
T2V (int.)	91.7	97.7	97.7	96.8	64.7	**54.7**	59.2	**86.9**
T2V (ext.)	91.3	97.5	97.8	96.9	58.6	54.6	**60.0**	86.1
T2V 2x (int.)	91.0	97.3	97.2	**97.0**	**70.3**	54.1	59.4	85.8
Initial Method	**Multiple Objects**	**Human Action**	**Color**	**Spatial Relationship**	**Scene**	**App. Style**	**Temp. Style**	**Overall Cons.**
T2I	**30.8**	92.2	85.0	**29.9**	45.2	21.1	25.0	26.0
T2V (int.)	25.5	**95.4**	85.3	28.6	41.4	**21.2**	25.3	26.6
T2V (ext.)	28.5	95.2	84.2	25.9	36.8	20.9	25.6	**26.8**
T2V 2x (int.)	29.3	94.0	**87.7**	28.6	**44.2**	20.9	**25.7**	26.7

Table 18. Detailed VBench metrics of different model initialization methods for higher frame count T2V model training.

Model	Source	PC	SA	Avg.
OpenSora [70]	Open	35	21	28
SVD [3]	Open	34	37	35
CogVideoX-2B [64]	Open	39	40	39
LaVIE [59]	Open	36	45	41
VideoCrafter2 [7]	Open	36	47	41
CogVideoX-5B [64]	Open	41	57	49
Model	**Source**	**PC**	**SA**	**Avg.**
Gen-2 [51]	Closed	31	26	29
Pika [44]	Closed	33	25	29
Lumiere-T2V [2]	Closed	31	35	33
Lumiere-T2I2V [2]	Closed	25	46	35
Luma Dream Machine [39]	Closed	30	53	41.5
XL	Our	36	57	47
M-512	Our	**43**	**59**	**51**

Table 19. Performance of T2V models on VideoPhy [1].

H. Flexible Applications

Here, we demonstrate how to extend our STIV to various applications, such as video prediction, frame interpolation, multi-view generation, and long video generation.

Method	IS ↑	FVD ↓	VBench-Quality ↑
CogVideo [29]	50.5	626	-
TATS [21]	79.3	332	-
MMVG [19]	73.7	328	-
VideoFusion [40]	80.0	173	-
MAGVIT [65]	83.6	159	-
XL-128	<u>93.4</u>	124	69.9
- Spatial Mask	88.5	**102**	**70.6**
+ Temporal Mask	**94.9**	167	68.1
+ Temporal ScaleShiftGate	78.9	141	69.1
+ Causal TemporalAttention	86.9	<u>106</u>	<u>70.3</u>

Table 20. Performance of Class-to-Video Generation on UCF-101.

Video Prediction We initialize from a STIV-XXL model to train a text-video-to-video model conditioned on the first four frames. As shown in Fig. 21a, the video-to-video model (STIV-V2V) shows significantly lower FVD scores compared to the text-to-video model (T2V) on MSRVTT [61] test set and MovieGen Bench [46]. This result indicates that video-to-video models can achieve superior performance, which is promising for applications in

Model	MSRVTT FVD ↓	MovieGen FVD ↓
T2V	536.2	347.2
STIV-V2V	**183.7**	**186.3**

(a) Comparison of T2V and V2V.

Model	use text	MSRVTT FID ↓	MSRVTT FVD ↓
STIV-TUP	No	2.2	6.3
STIV-TUP	Yes	**2.0**	**5.9**

(b) Performance of STIV-TUP.

Model	GSO [17] PSNR ↑	GSO [17] SSIM ↑	GSO [17] LPIPS ↓
Zero123++	21.200	0.723	**0.143**
STIV-TI2V-XL	**21.643**	**0.724**	0.156

(c) Multiview generation comparison.

autonomous driving and embodied AI where high fidelity and consistency in generated video frames are crucial.

Frame Interpolation We propose STIV-TUP, a temporal upsampler initialized from an STIV-XL model, and continue train conditioned on consecutive frames sampled by stride of 2 with the text conditioning. Fig. 21b shows that our STIV can also be used to do decent frame interpolation conditioned on both text and image. We observe that using text conditions is slightly better in FID and FVD on the MSRVTT test set. We also cascade the temporal upsampler with our main model to explore whether it can boost the main performance. As shown in Tab. 8 and Tab. 3, using a temporal upsampler on the top of the main models can improve the quality performance while maintaining other scores.

Multi-View Generation Multi-view generation is a specialized task focused on creating novel views from a given input image. This task places demands on view consistency and can greatly benefit from a well-pretrained video generation model. By adapting a video generation model for multi-view generation, we can evaluate whether the pretraining has effectively captured underlying 3D information, which would enhance multi-view generation.

Here, we adopt the novel view camera definitions outlined in Zero123++ [55], which specifies six novel view cameras for each input image. The initial frame in our TI2V model is set as the given image, and the next six frames, representing novel views, are predicted as future frames within TI2V. For training, we began with our TI2V-XL checkpoint trained with a 256 resolution, fine-tuning it for 110k steps on Objaverse [15]. For a fair comparison, we increased the image resolution to 320 during fine-tuning, aligning with the settings used in Zero123++. Our evaluation used objects from the Google Scanned Objects dataset [17], where we compared the output multi-view images against ground-truth renderings. As shown in Fig. 21c, despite only using temporal attention for cross-view consistency, our approach achieves comparable performance to Zero123++ which uses full attention to all the views. This outcome validates the effectiveness of our spatiotemporal attention in maintaining 3D consistency. A visual comparison between our approach and Zero123++ is shown in Fig. 7.

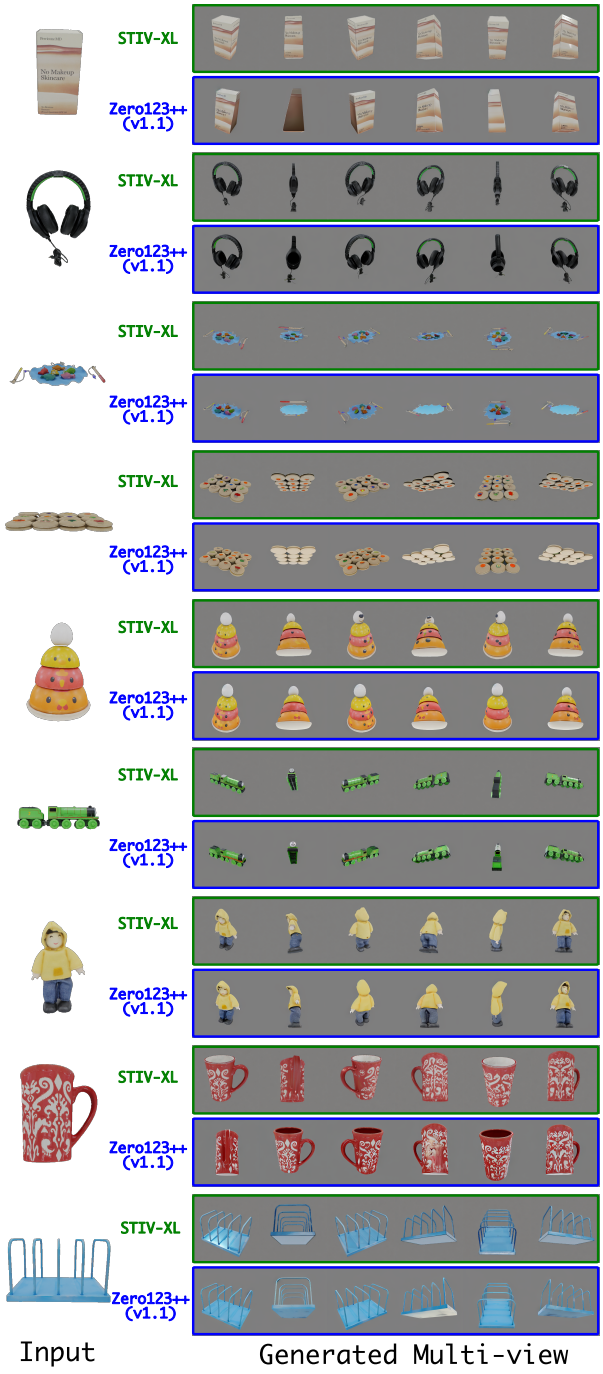


Figure 7. The visual comparison between our STIV-XL with Zero123++ [55] on GSO [17].

I. Detailed Results for Imaging Dropout

As mentioned in Section 3.2.2, after adding imaging dropout. We observe this phenomenon happens when we scale our model to 8B with 512 or higher resolutions, probably due to the model being more easily overfitting to follow the first frame with a larger model capacity, and it becomes worse under the higher resolution. Specifically, we showcase some examples to see the different between generated videos without image dropout and videos with image dropout (STIV-M-512). We generate the videos conditioned on the first frame and text prompt borrowed from MovieGenBench [46] As shown in Fig. 8 to 10, using image condition dropout in general achieves better performance than the baseline in terms of motion quality.

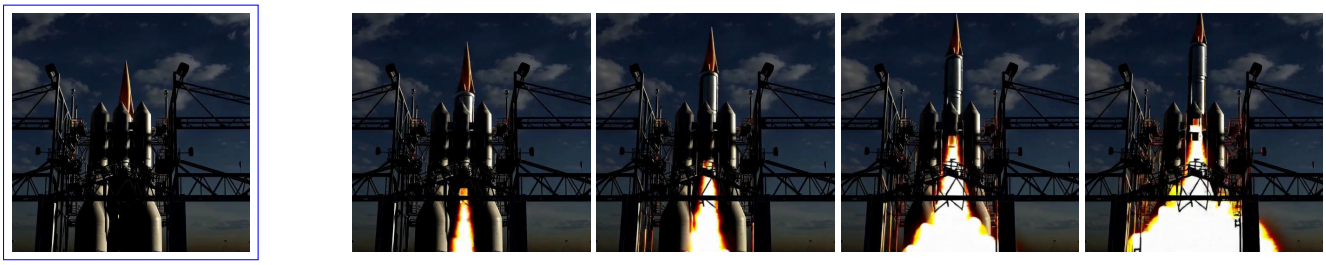
TI2V-M-512 V.S. STIV-M-512

Prompt: *A red panda taking a bite of a pizza.*

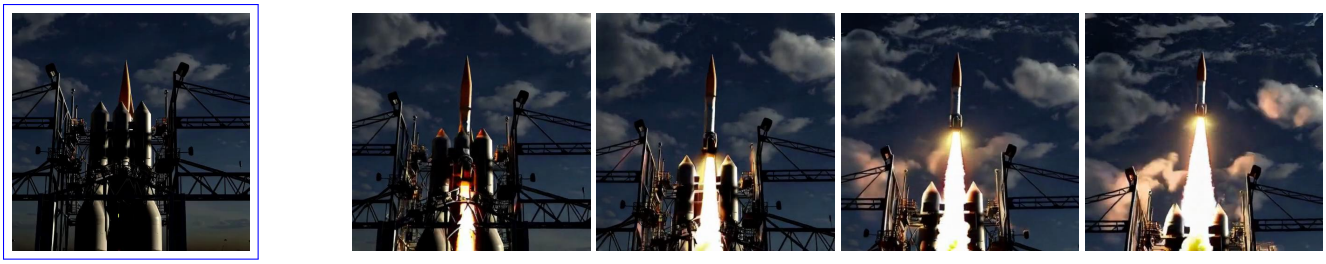
Reference Image

Reference Image

Prompt: *A rocket blasting off from the launch pad, accelerating rapidly into the sky.*



Reference Image



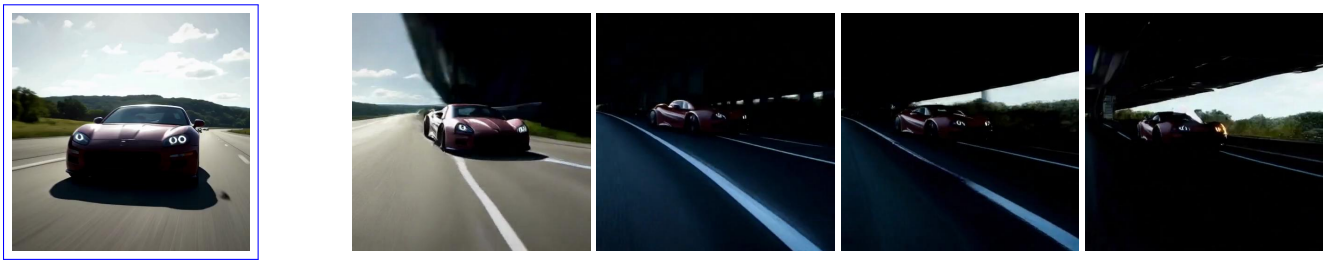
Reference Image

Figure 8. Visualization of TI2V-M-512 V.S. STIV-M-512. (Given the same prompt, the figures in the top row are generated by TI2V-M-512, while the figures in the bottom row are generated by STIV-M-512.)

TI2V-M-512 V.S. STIV-M-512

Prompt: *A sports car accelerating rapidly on an open highway, the engine roaring.*

Reference Image



Reference Image

Prompt: *A glass of iced coffee condensing water on the outside, with droplets forming and sliding down the glass in slow motion.*

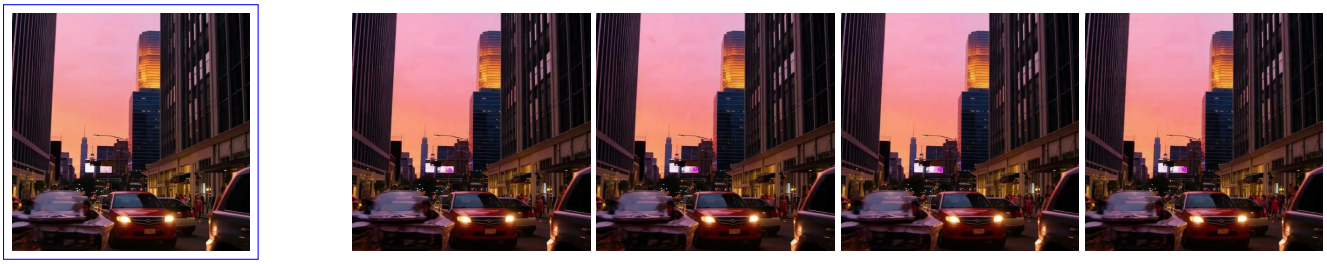
Reference Image

Reference Image

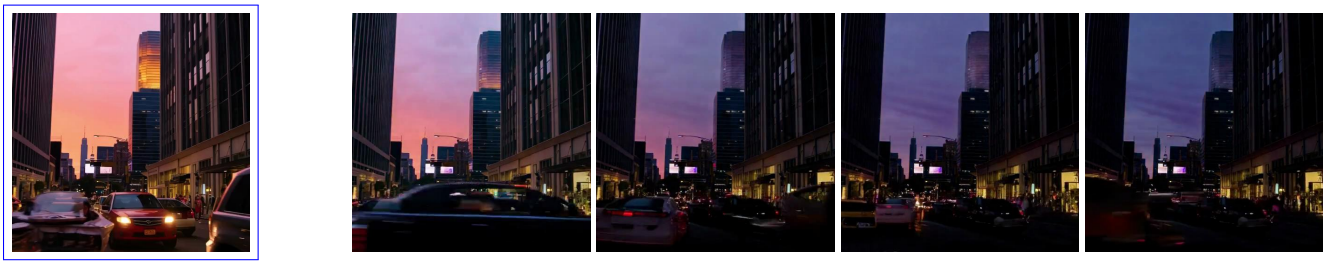
Figure 9. Visualization of TI2V-M-512 V.S. STIV-M-512. (Given the same prompt, the figures in the top row are generated by TI2V-M-512, while the figures in the bottom row are generated by STIV-M-512.)

TI2V-M-512 V.S. STIV-M-512

Prompt: *Cars and pedestrians move through a bustling downtown street lined with skyscrapers, their lights reflecting off the windows of the towering buildings as day turns to dusk.*

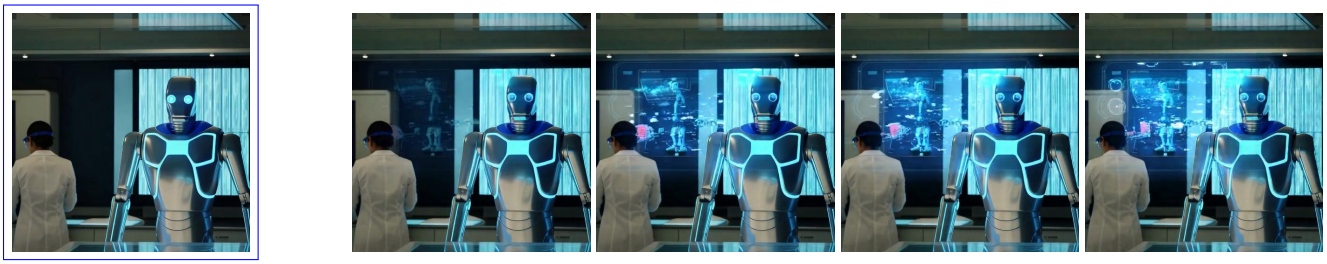


Reference Image

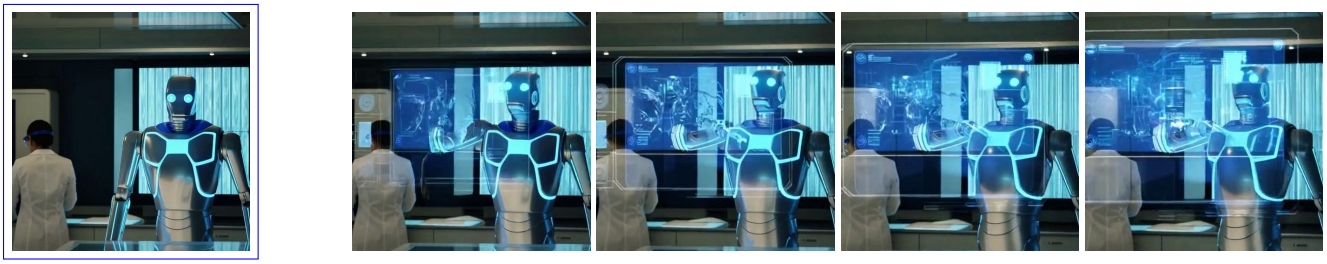


Reference Image

Prompt: *Robots move efficiently through a futuristic laboratory, adjusting holographic displays and conducting experiments, while scientists observe and interact with the high-tech equipment.*



Reference Image



Reference Image

Figure 10. Visualization of TI2V-M-512 V.S. STIV-M-512. (Given the same prompt, the figures in the top row are generated by TI2V-M-512, while the figures in the bottom row are generated by STIV-M-512.)

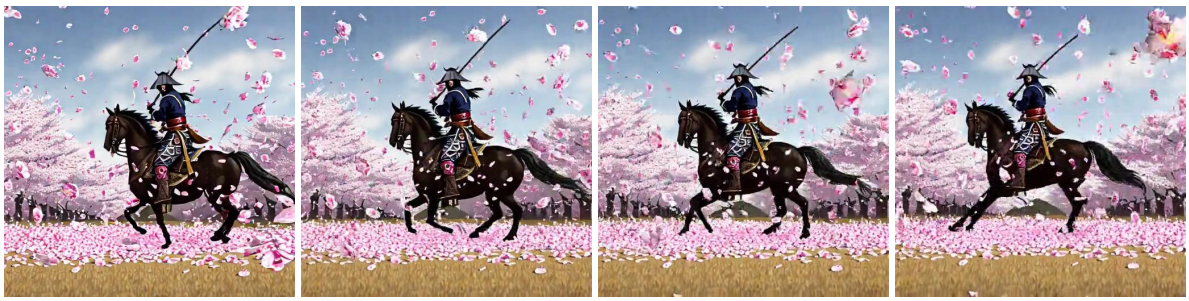
J. More Examples

We show more examples at the end of the Appendix using the text prompts and image as first frame condition borrowed from MovieGenBench [46] and Sora [42].

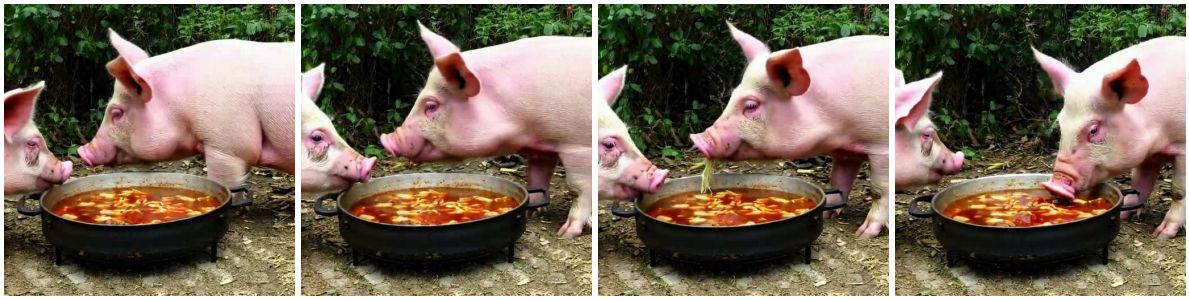
Text-to-Video

Prompt: *A pirate ship sailing through a storm with enormous waves crashing against the sides, its crew fighting against the wind as lightning illuminates the scene.*

Prompt: *A samurai on horseback charging across a field of cherry blossoms, slicing petals in mid-air as they fall, leaving a trail of pink in their path.*



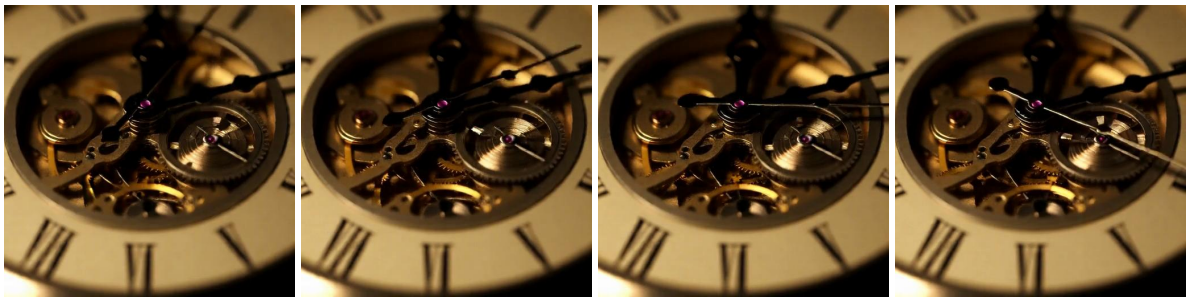
Prompt: *Two pigs are eating a hotpot.*



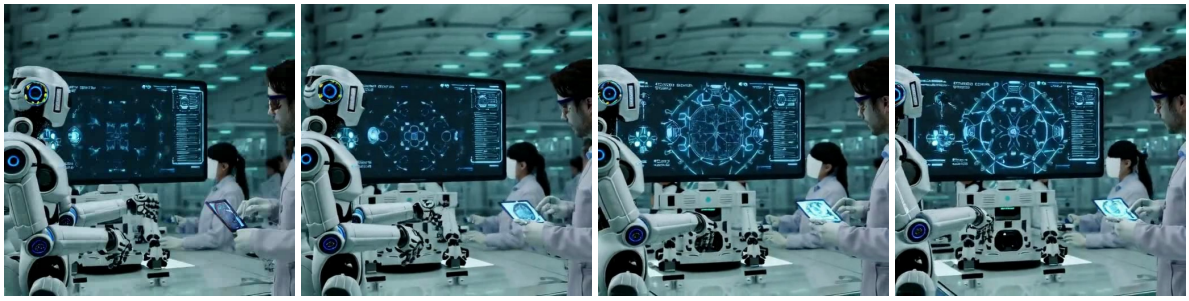
Prompt: *Giant Pandas are eating hot noodles in a Chinese restaurant.*

Text-to-Video

Prompt: *A zoom-in on a clock face, focusing on the intricate movement of the hands and the ticking mechanism inside.*



Prompt: *Robots move efficiently through a futuristic laboratory, adjusting holographic displays and conducting experiments, while scientists observe and interact with the high-tech equipment.*



Prompt: *A robotic arm wielding a glowing sword, battling a shadowy figure in a high-tech dojo, each strike creating sparks that light up the space.*

Prompt: *A city skyline reflected in the water, but the reflection shows an alternate world with flying cars, towering robots, and futuristic architecture.*

Text-to-Video

Prompt: *A dog dressed as a chef, expertly flipping pancakes in a kitchen.*

Prompt: *A motocross bike accelerating out of a tight turn on a dirt track.*

Prompt: *A snowboarder performing a dramatic backflip over a frozen lake, landing gracefully and leaving a trail of sparkling ice dust in the air.*

Prompt: *A surfer accelerating on a wave, carving through the water.*

Text-to-Video

Prompt: *A person dancing with their own shadow, which has come to life.*

Prompt: *A bobsled team accelerating down an icy track.*

Prompt: *A cyclist accelerating out of the saddle during a steep climb.*

Prompt: *A speed skater accelerating during a short track race.*

Text-Image-to-Video

Prompt: *Reflections in the window of a train traveling through the Tokyo suburbs.*

Reference Image

Prompt: *The Glenfinnan Viaduct is a historic railway bridge in Scotland, UK, that crosses over the west highland line between the towns of Mallaig and Fort William. It is a stunning sight as a steam train leaves the bridge...*

Reference Image

Prompt: *The camera follows behind a white vintage SUV with a black roof rack as it speeds up a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up from it's tires, the sunlight shines on the SUV...*

Reference Image

Prompt: *Photorealistic closeup video of two pirate ships battling each other as they sail inside a cup of coffee.*

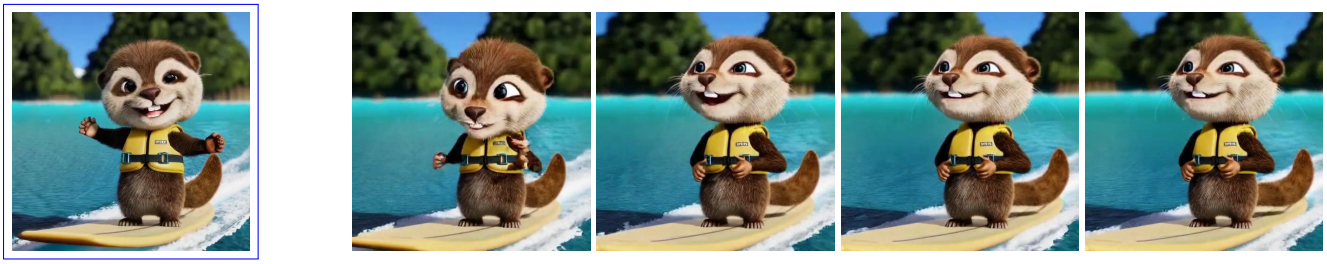
Reference Image

Text-Image-to-Video

Prompt: *A litter of golden retriever puppies playing in the snow. Their heads pop out of the snow, covered in.*

Reference Image

Prompt: *An adorable happy otter confidently stands on a surfboard wearing a yellow lifejacket, riding along turquoise tropical waters near lush tropical islands, 3D digital render art style.*



Reference Image

Prompt: *A dog dressed as a chef, expertly flipping pancakes in a kitchen.*

Reference Image

Prompt: *A skeleton wearing a flower hat and sunglasses dances in the wild at sunset.*

Reference Image

Text-Image-to-Video

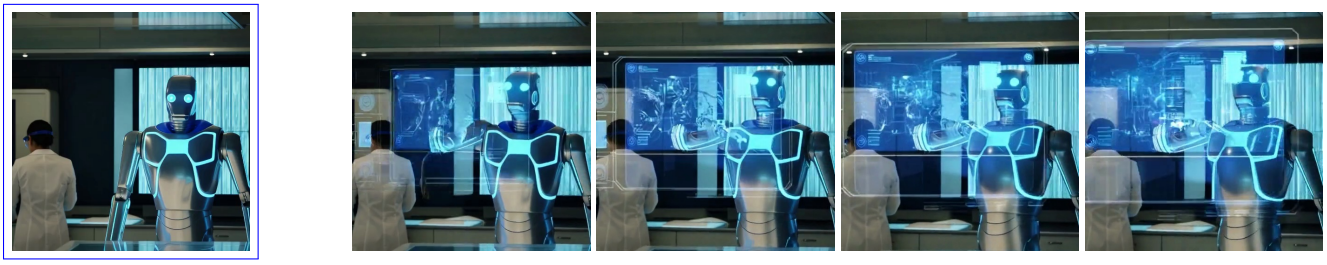
Prompt: *The video features a central spacecraft with a predominantly white and gray color scheme, accented with red and black details. It has a sleek, angular design with multiple protruding elements that suggest advanced technology...*

Reference Image

Prompt: *The video begins with a dark space background, dotted with stars, and a central object that appears to be a spacecraft with a glowing blue light at its core. The spacecraft is detailed with various components...*

Reference Image

Prompt: *Robots move efficiently through a futuristic laboratory, adjusting holographic displays and conducting experiments, while scientists observe and interact with the high-tech equipment.*



Reference Image

Prompt: *The video presents a serene scene with a group of camels walking in a line across a desert landscape. The camels are adorned with colorful saddles and are led by a person wearing a green garment. The background features a clear sky...*

Reference Image

Text-Image-to-Video

Prompt: *A crab made of different jewlery is walking on the beach. As it walks, it drops different jewelry pieces like diamonds, pearls, etc.*

Reference Image

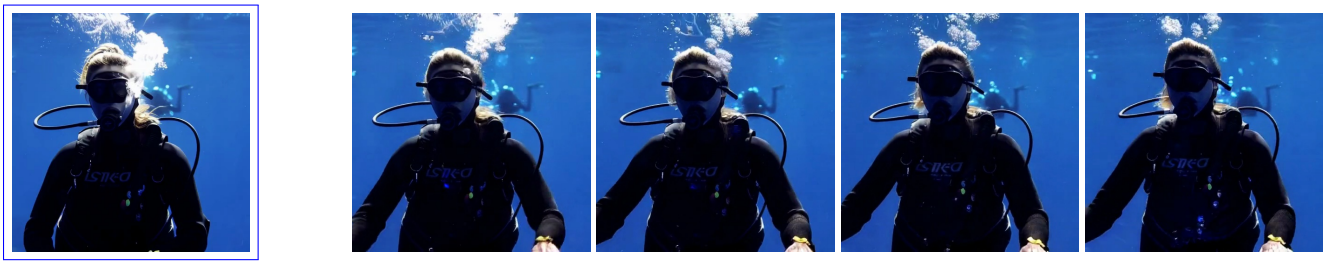
Prompt: *The video captures a single sea turtle with a patterned shell and flippers, swimming in a clear blue underwater environment. The turtle moves gracefully over a bed of coral reefs, which exhibit a variety of colors...*

Reference Image

Prompt: *A mesmerizing video of a jellyfish moving through water, with its tentacles flowing gracefully.*

Reference Image

Prompt: *A video of a diver creating bubbles underwater, with bubbles rising and interacting with each other.*



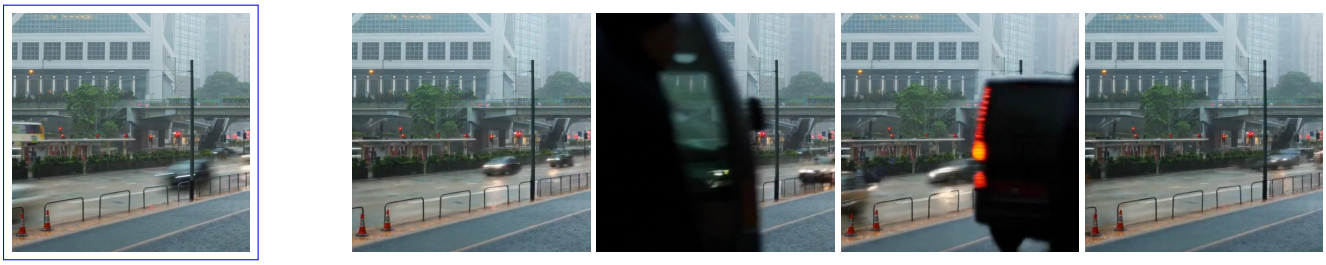
Reference Image

Text-Image-to-Video

Prompt: *The individual in the video is dressed in a blue protective suit with a hood, a mask with a filter, and white gloves. They are holding a spray bottle in one hand and a spray nozzle in the other...*

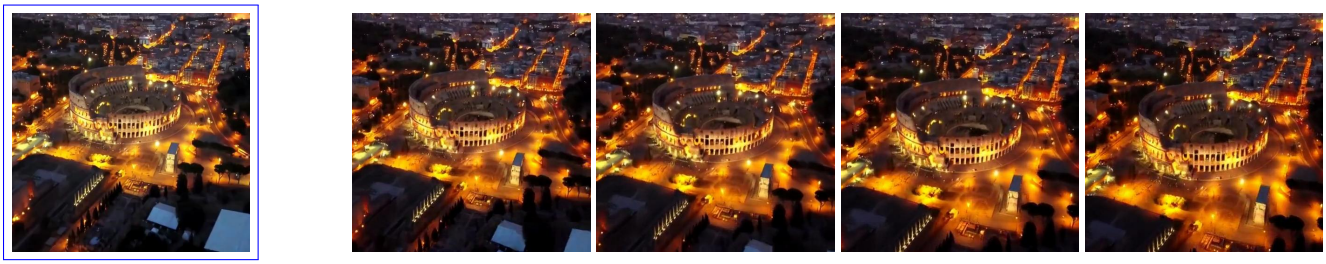
Reference Image

Prompt: *The video captures a bustling city street scene during the evening. The sky is overcast, and the street is wet, reflecting the lights from the vehicles and buildings. The buildings are tall with modern architecture...*



Reference Image

Prompt: *The video presents a series of images capturing the Colosseum from an aerial perspective during the evening. The ancient amphitheater is illuminated by artificial lighting, which highlights its circular shape and the arches...*



Reference Image

Prompt: *The video features two dogs, one with a predominantly white coat and the other with a mix of black, brown, and white fur. Both dogs are adorned with accessories; the white dog wears a red tie, while the other sports a purple bow tie...*

Reference Image