STIV: Scalable Text and Image Conditioned Video Generation

Zongyu Lin[*1] **Wei Liu**[*1] **Chen Chen**[*2] **Jiasen Lu**[*2] **Wenze Hu**[*2]

Tsu-Jui Fu[*2] Jesse Allardice[*2] Zhengfeng Lai[*2] Liangchen Song[*2] Bowen Zhang[*2]

Cha Chen[*2] Yiran Fei[*] Yifan Jiang[*] Lezhi Li[*] Yizhou Sun[◇3]

Kai-Wei Chang[◇3] Yinfei Yang[*]

[*]**Apple** [◇]**UCLA**

Abstract

The field of video generation has made remarkable advancements, yet there remains a pressing need for a clear, systematic recipe that can guide the development of robust and scalable models. In this work, we present a comprehensive study that systematically explores the interplay of model architectures, training recipes, and data curation strategies, culminating in a simple and scalable text-image-conditioned video generation method, named STIV. Our framework integrates image condition into a Diffusion Transformer (DiT) through frame replacement, while incorporating text conditioning via a joint image-text conditional classifier-free guidance. This design enables STIV to perform both text-to-video (T2V) and text-image-to-video (TI2V) tasks simultaneously. Additionally, STIV can be easily extended to various applications, such as video prediction, frame interpolation, multi-view generation, and long video generation, etc. With comprehensive ablation studies on T2I, T2V, and TI2V, STIV demonstrate strong performance, despite its simple design. An 8.7B model with 512^2 resolution achieves 83.1 on VBench T2V, surpassing both leading open and closed-source models like CogVideoX-5B, Pika, Kling, and Gen-3. The same-sized model also achieves a state-of-the-art result of 90.1 on VBench I2V task at 512^2 resolution. By providing a transparent and extensible recipe for building cutting-edge video generation models, we aim to empower future research and accelerate progress toward more versatile and reliable video generation solutions.

1 Introduction

The field of video generation has witnessed a significant progress with the introduction of Sora [43], a video generation model based on Diffusion Transformer (DiT) [44] architecture. Researchers have been actively exploring optimal methods to incorporate text and other conditions into the DiT architecture. For example, PixArt-α [10] leverages cross attention, while SD3 [20] concatenates text with the noised patches and applies self-attention using the MMDiT block. Several video generation models [65, 47, 22] adopt similar approaches and have made substantial progress in the text-to-video (T2V) task. Pure T2V approaches often struggle with producing coherent and realistic videos, as their outputs are not grounded in external references or contextual constraints [14]. To address this limitation, text-image-to-video (TI2V) introduce an initial image frame along with the textual prompt, providing a more concrete grounding for the generated video.

Despite substantial progress in video generation, achieving Sora-level performance for T2V and TI2V remains challenging. A central challenge is how to seamlessly integrate image-based conditions into the DiT architecture, calling for innovative techniques blend visual inputs smoothly with textual cues. Meanwhile, there is a pressing need for stable, efficient large-scale training strategies, as well as improving the overall quality of training datasets. To address these issues, a comprehensive, step-by-step "recipe" would greatly assist in developing unified models that handle both T2V and TI2V task under one framework. Overcoming these challenges is essential for advancing the field and fully realizing the potential of video generation models.

Although various studies [62, 4, 8, 12, 50, 70, 15] have examined methods of integrating image conditions into the U-Net architectures, how to effectively incorporate such conditions into the DiT
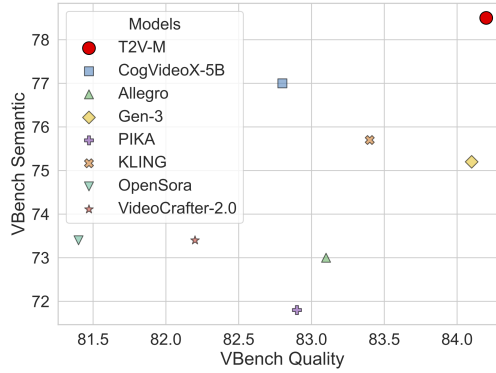
Figure 1: Performance comparison of our Text-to-Video model against both open-source and closed-source state-of-the-art models on VBench [32].

architecture remains unsolved. Moreover, existing studies in video generation often focuses on individual aspects independently, overlooking the how their collective impact on overall performance. For instance, while stability tricks like QK-norm [29, 20] have been introduced, they prove insufficient as models scale to larger sizes [58], and no existing approach has successfully unified T2V and TI2V capabilities within a single model. This lack of systematic, holistic research limits progress toward more efficient and versatile video generation solutions.

In this work, we first present a comprehensive study of model architectures and training strategies to establish a robust foundation for T2V. Our analysis reveals three key insights: (1) stability techniques such as QK-norm and sandwich-norm [26, 18] are critical for effectively scaling larger video generation models; (2) employing factorized spatial-temporal attention [3], MaskDiT [71], and switching to AdaFactor [55] significantly improve training efficiency and reduce memory usage with minimal impact on performance loss; (3) progressive training, where spatial and temporal layers are initialized from separate models, outperforms using a single model under the same compute constraints. Starting from a PixArt-α baseline architecture, we address scaling challenges with these stability and efficiency measures, and further enhance performance with Flow Matching [42], RoPE [57], and micro conditions [46]. Our largest T2V model (8.7B parameters) achieves state-of-the-art semantic alignment and a VBench score of 83.1.

We then identify the optimal model architecture and hyperparameters established in the T2V setting and apply them to the TI2V task. Our results show that simply replacing the first noised latent frame with the un-noised image condition latent yields strong performance. Although ConsistI2V [50] introduced a similar idea in a U-Net setting, it required spatial self-attention for each frame and window-based temporal self-attention to match our quality. In contrast, the DiT architecture natively propagates the image-conditioned first frame through stacked spatial-temporal attention layers, eliminating the need for these additional operations. However, as we scale up spatial resolution, we observe the model producing slow or nearly static motion. To solve this, we introduce random dropout of the image condition during training and apply joint image-text conditional classifier-free guidance (JIT-CFG) for both text and image conditions during inference. This strategy resolves the motion issue and also enables a single model to excel at both T2V and TI2V tasks.

With all these changes, we finalize our model and scale it up from 600M to 8.7B parameters. Our best STIV model achieves a state-of-the-art result of 90.1 in the VBench I2V task at 512^2 resolution. Beyond enhancing video generation quality, we demonstrate the potential of extending our framework to various downstream applications, including video prediction, frame interpolation, multi-view generation and long video generation. These results validate the scalability and versatility of our approach, showcasing its ability to address diverse video generation challenges. We summarize our contributions as follows:

- We present STIV, a single model capable of performing both T2V and TI2V tasks. At its core, we replace the noised latent with the un-noised image condition latent and introduce joint image-text conditioned CFG.
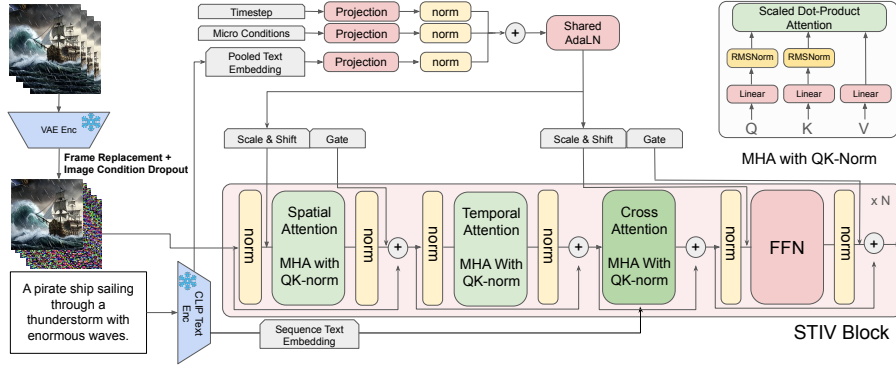
Figure 2: We replace the first frame of the noised video latents with the ground truth latent and randomly drop out the image condition. We use cross attention to incorporate the text embedding, and use QK-norm in multi-head attention, the sandwich-norm in both attention and feedforward, and stateless layernorm after singleton conditions to stabilize the training.

- We conduct a systematic study for T2I, T2V and TI2V, covering model architectures, efficient and stable training techniques, and progressive training recipes to scale up the model size, spatial resolution, and duration.

- These design features make it easy to train and adaptable to various tasks, including video prediction, frame interpolation, and long video generation.

- We include detailed ablation studies on different design choices and hyperparameters, evaluated on VBench, VBench-I2V and MSRVTT. These studies demonstrate the effectiveness of the proposed model compared with a range of recent state-of-the-art open-source and closed-source video generation models.

2 STIV

This section describes our key components of our proposed STIV method for text-image-to-video (TI2V) generation, which is illustrated in Fig. 2. Afterward, Sec. A and B presents detailed experimental results.

2.1 BASE MODEL ARCHITECTURE

The STIV model is based on PixArt-α [10], which converts the input frames into spatial and temporal latent embeddings using a frozen Variational Autoencoder (VAE). These embeddings are then processed by a stack of learnable DiT-like blocks. We employ the T5 [49] tokenizer and an internally trained CLIP [48] text encoder to process text prompts. The overall framework is illustrated in Fig. 2. Other significant architectural changes are outlined below.

Spatial-Temporal Attention We use factorized spatial and temporal attention [3]. We first fold the temporal dimension into the batch dimension and perform spatial self-attention on spatial tokens. Then, we permute the outputs and fold the spatial dimension into the batch dimension to perform temporal self-attention on temporal tokens. By using factorized spatial and temporal attention, we can easily preload weights from a text-to-image (T2I) model, as images are a special case of videos with only one temporal token and only need spatial attention.

Singleton Condition We use the original image resolution, crop coordinates, and sampling stride as micro conditions to encode the meta information of the training data. We first use a sinusoidal embedding layer to encode these properties, followed by an MLP to project them into a d-dimensional embedding space. These micro condition embeddings, along with the diffusion timestep embedding and the last text token embedding from the last layer of the CLIP model, are added to form a singleton condition. We also apply stateless layer normalization to each singleton embedding and then add

them together. This singleton condition is used to produce shared scale-shift-gate parameters that are utilized in the spatial attention and feed-forward layers of each Transformer layer.

Rotary Positional Embedding Rotary Positional Embeddings (RoPE) [57] are used so that the model has a strong inductive bias for processing relative temporal and spatial relationships. Additionally, RoPE can be made compatible with the masking methods used in high compute applications and are highly adaptable to variations in resolution [74]. We apply 2D RoPE [39] for the spatial attention and 1D RoPE for the temporal attention inside the factorized Spatial-Temporal attention.

Flow Matching Instead of using the conventional diffusion loss, we opt for a Flow Matching training objective. It defines a conditional optimal transport between two examples drawn from a source and target distribution. In our case, we assume the source distribution to be Gaussian and utilize linear interpolates [42] to achieve this.

$$\boldsymbol{x}_t = t \cdot \boldsymbol{x}_1 + (1-t) \cdot \boldsymbol{\epsilon}. \tag{1}$$

The training objective is then formulated as

$$\min_\theta \mathbb{E}_{\boldsymbol{x}, \boldsymbol{\epsilon} \in \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \boldsymbol{c}, t} \left[\|\boldsymbol{F}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t) - \boldsymbol{v}_t\|_2^2 \right] \tag{2}$$

where the velocity vector field $\boldsymbol{v}_t = \boldsymbol{x_1} - \boldsymbol{\epsilon}$.

In inference time, we solve the corresponding reverse-time SDE, from timestep 0 to 1, to generate images from a randomly sampled Gaussian noise $\boldsymbol{\epsilon}$.

2.2 MODEL SCALING

As we scale up the model, we encounter training instability and infrastructure challenges. In this section, we outline the methods to improve training stability and efficiency.

Stable Training Recipes We discovered that QK-Norm — applying RMSNorm [68] to the query and key vectors prior to computing attention logits — significantly stabilizes training. This finding aligns with the results reported in SD3 [20]. We also change from pre-norm to sandwich-norm [18] for both MHA and FFN, which involves adding pre-norm and post-norm with stateless layer normalization [37] to both the layers within the STIV block.

$$\text{MHA}(x) = x + \text{gate} \cdot \text{norm} \left(\text{Attn} \left(\text{scale} \cdot \text{norm}(x) + \text{shift} \right) \right)$$
$$\text{FFN}(x) = x + \text{gate} \cdot \text{norm} \left(\text{MLP} \left(\text{scale} \cdot \text{norm}(x) + \text{shift} \right) \right)$$

Efficient DiT Training We follow MaskDiT [71] by randomly masking 50% of spatial tokens before passing them into the major DiT blocks. After unmasking, we add two additional DiT blocks. We also switch from AdamW to AdaFactor optimizer and employ gradient checkpointing to only store the self-attention outputs. These modifications significantly enhance efficiency and reduce memory consumption, enabling the training of larger models at higher resolution and longer duration.

2.3 IMAGE CONDITIONING

2.3.1 FRAME REPLACEMENT

During training, we replace the noised first frame latent with the un-noised latent of the image condition before passing the latents into the STIV blocks, and masking out the loss of the replaced frame. During inference, we use the un-noised latent of the original image condition for the first frame at each TI2V diffusion step.

The frame replacement strategy offers flexibility in extending STIV to various applications. For instance, if $\boldsymbol{c}_I = \varnothing$, it defaults to text-to-video (T2V) generation. Conversely, if $\boldsymbol{c}_I$ is the initial frame, it becomes the typical text-image-to-video (TI2V) generation. Moreover, if multiple frames as $\boldsymbol{c}_I$ are provided, they can be used for video prediction even without $\boldsymbol{c}_T$. Additionally, supplying the first and last frames as $\boldsymbol{c}_I$ enables the model to learn a frame interpolation, generating frames between them. Furthermore, combining T2V and frame interpolation allows for the generation of long-duration videos: T2V generates keyframes, and frame interpolation frames then fills in frames

between each pair of consecutive keyframes. Ultimately, a single model can be trained to perform all tasks by randomly selecting the appropriate conditioning strategy.

2.3.2 IMAGE CONDITION DROPOUT

As mentioned above, the frame replacement strategy offers substantial flexibility for training various types of models. Here, we demonstrate a specific application in which we train a model to perform both T2V and TI2V tasks. In this case, we randomly drop out c_I and c_T during training.

Classifier-free guidance (CFG), commonly used in text-to-image generation, has proven to be highly beneficial in enhancing the quality of generated images by directing the probability mass toward the high-likelihood regions given the condition. Building on this concept, we introduce a **Joint Image-Text Classifier-Free Guidance (JIT-CFG)** approach, which leverages both text and image conditions. It modifies the velocity estimates as

$$\hat{F}_\theta(x_t, c_T, c_I, t) = F_\theta(x_t, \varnothing, \varnothing, t) \\ + s \cdot (F_\theta(x_t, c_T, c_I, t) - F_\theta(x_t, \varnothing, \varnothing, t))$$

where s is the guidance scale. When $c_I = \varnothing$, it reduces to standard CFG for T2V generation. Although it is possible to introduce two separate guidance scales, as done in [6], to balance the strength of the image and text conditions, we found that our two-pass approach yields strong results. Additionally, using two scales would require three forward passes, increasing the inference cost.

Empirical observations A.4.2 suggest that applying image condition dropout with JIT-CFG effectively not only achieves multi-task training in a natural way, but also resolves the staleness issue for a 512^2 STIV model. We hypothesize that image condition dropout prevents the model from passively overfitting to the image condition, allowing it to more effectively capture the motion information from the underlying video training data.

2.4 PROGRESSIVE TRAINING RECIPE

We adopt a progressive training recipe as illustrated in Figure 3. The process begins by training a text-to-image (T2I) model, which serves to initialize a text-to-video (T2V) model. Next, the T2V model serves as the starting point for initializing the STIV model. To facilitate rapid adaptation to higher resolutions and longer durations training, we incorporate interpolated RoPE embeddings in both the spatial and temporal dimensions, and initialize the model weights using those from the lower-resolution, shorter-duration models.
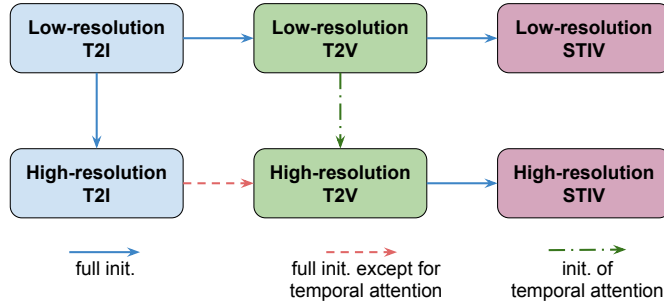

Figure 3: Progressive training pipeline of the STIV model. The T2I model is first trained to initialize the T2V model, which then initializes the STIV model at both low and high resolutions. Notably, the high-res T2V model is initialized using both the high-res T2I model and the low-res T2V model.

3 CONCLUSION

In conclusion, we conduct a comprehensive study on how to build a good video generation model, and present a scalable and flexible approach for integrating text and image conditioning within a unified video generation framework. Our model not only demonstrates good performance on public

benchmarks, but also shows versatility in downstream applications, supporting controllable video generation, video prediction, frame interpolation, long video generation, and multi-view generation, which collectively highlight its potential as a foundation for the broad research community.

REFERENCES

[1] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, pp. 4, 2021.

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023.

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

[7] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[9] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.

[10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[11] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70M: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024.

[12] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023.

[13] Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024.

[14] Zhixuan Chu, Lei Zhang, Yichen Sun, Siqiao Xue, Zhibo Wang, Zhan Qin, and Kui Ren. Sora detector: A unified hallucination detection for large text-to-video models. *arXiv preprint arXiv:2405.04180*, 2024.

[15] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance. *arXiv e-prints*, pp. arXiv–2311, 2023.

[16] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Fine-grained open domain image animation with motion guidance. *arXiv preprint arXiv:2311.12886*, 2023.

[17] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

[18] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

[19] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

[20] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[21] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell Me What Happened: Unifying Text-guided Video Completion via Multimodal Masked Video Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[22] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.

[23] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic VQGAN and time-sensitive transformer. In *European Conference on Computer Vision*, pp. 102–118, 2022.

[24] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. In *European Conference on Computer Vision*, 2023.

[26] Xinyu Gong, Wuyang Chen, Tianlong Chen, and Zhangyang Wang. Sandwich Batch Normalization: A Drop-In Replacement for Feature Distribution Heterogeneity. In *Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[27] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[28] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

[29] Alex Henry, Prudhvi Raj Dachapally, Shubham, and Yuxuan Chen. Query-Key Normalization for Transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL `https://arxiv.org/abs/1706.08500`.

[31] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[32] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

[33] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijaya-narasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. In *arXiv:1412.0767*, 2014.

[34] Kuaishou. Kling. `https://kling.kuaishou.com`, 2024.

[35] Zhengfeng Lai, Vasileios Saveris, Chen Chen, Hong-You Chen, Haotian Zhang, Bowen Zhang, Juan Lao Tebar, Wenze Hu, Zhe Gan, Peter Grasch, et al. Revisit large-scale image-caption data in pre-training multimodal foundation models. *arXiv preprint arXiv:2410.02740*, 2024.

[36] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pp. 111–127. Springer, 2025.

[37] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pp. arXiv–1607, 2016.

[38] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2V-Turbo: Breaking the Quality Bottleneck of Video Consistency Model with Mixed Reward Feedback. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[39] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024.

[40] Luma. Dream machine. `https://lumalabs.ai/dream-machine`, 2024.

[41] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[42] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024.

[43] OpenAI. Sora. `https://openai.com/index/sora`, 2024.

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

[45] Pika. Pika 1.0. `https://pika.art`, 2023.

[46] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[47] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[49] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[50] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[52] Runway. Gen-2. `https://research.runwayml.com/gen2`, 2023.

[53] Runway. Gen-3 alpha. `https://runwayml.com/research/introducing-gen-3-alpha`, 2024.

[54] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[55] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

[56] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

[57] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[58] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[59] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *International Conference on Computer Vision (ICCV)*, 2015.

[60] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[61] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[62] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

[63] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023.

[64] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5288–5296, 2016.

[65] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *arXiv:2408.06072*, 2024.

[66] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, JiaZheng Xu, Yuanming Yang, Xiaohan Zhang, Xiaotao Gu, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[67] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, and Lu Jiang Irfan Essa. MAGVIT: Masked Generative Video Transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[68] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[69] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.

[70] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

[71] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.

[72] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing Efficient Video Production for All, 2024. URL `https://github.com/hpcaitech/Open-Sora`.

[73] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024.

[74] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024.

A RECIPE STUDY FOR STIV

A.1 BASIC SETUP

We first introduce the training, data and evaluation setup before diving into our model and studies as follows.

Training Unless otherwise specified, we use the AdaFactor optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) [55] without any weight decay. We also clip the gradient norm if the gradient norm exceeds 1.0. We use a constant learning rate schedule with a 1k step linear warmup with a maximum learning rate of 2×10^{-4}. For T2I models, we train each model for 400k steps with a batch size of 4,096. This is approximately 1.4 epochs on our internal T2I datasets. For T2V and TI2V models, we train each model for 400k steps with a batch size of 1,024. This is roughly 5.5 epochs on our internal video datasets. For all models, exponential moving average weights are gathered by a decay rate of 0.9999 and are then used for evaluation. When MaskDiT is used, we train with 50% spatial random masking during the initial 400k steps. Subsequently, we perform unmasked fine-tuning using all tokens. We use 50k steps of unmasked fine-tuning for T2I models and 100k steps for T2V and TI2V models.

Data We build a video data engine pipeline that includes video pre-processing, captioning, and filtering to accelerate the model's development when handling large-scale videos is required. Specifically, we apply PySceneDetect [1] to analyze video frames, detect and segment scenes based on abrupt transitions and gradual fades. This segmentation is followed by the feature extractions for filtering, including motion score, aesthetic score, text area, frame dimensions, clarity score, temporal consistency, and video orientation, among others. For each video segment, we perform dense captioning and categorization to gain a comprehensive understanding of video distribution. To further enhance caption quality, we adapt DSG [13] and propose DSG-Video, a metric designed to assess hallucination rates and overall caption quality. This data engine is integral in filtering videos and preparing tailored datasets for different training stages: our data sources include Panda-70M [11] and an internally curated high-quality dataset of 42M videos. Using this, we curate over 90M high-quality video-caption pairs. See more details in Appendix.

Evaluation We mainly use VBench [32], VBench-I2V and MSRVTT [64] to evaluate T2V and TI2V models. For VBench, we mainly report Quality (temporal quality and frame-wise quality), Semantic (semantic alignment with different perspectives of the input text prompt) and Total score (weighted average of Quality and Semantic), and they can be actually decomposed into 16 dimensions in total. 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. More details about the detailed dimensions are presented in Section I.1. We present three model scales: XL, XXL, and M with their configuration detailed in Table 1. In the following section, we use the notation X-S-T to represent an X-size model with an S^2 resolution and T frames. If unspecified, the default configuration is a 256^2 resolution with 20 frames. More detailed model and training configurations are provided in the Appendix.

Model Size	# of STIV Blocks	Hidden Dims	# of Attn Heads
XL (600M)	28	1,152	18
XXL (1.5B)	38	1,536	24
M (8.7B)	46	3,072	48

Table 1: Model Configurations

---

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

Model	COCO FID↓	Gen Eval↑	DSG Eval↑	Image Reward↑
Baseline	26.17	0.358	0.571	-0.25
+ QK norm	25.60	0.372	0.574	-0.22
+ Sandwich norm	25.76	0.366	0.577	-0.23
+ Cond. norm	25.58	0.393	0.583	-0.22
+ LR to 2E-4	26.35	0.379	0.586	-0.12
+ Flow	24.96	0.457	0.639	0.15
+ Renorm	21.16	0.471	0.668	0.32
+ AdaFactor	20.26	0.474	0.661	0.32
+ MaskDiT	23.85	0.499	0.663	0.30
+ Shared AdaLN	22.83	0.496	0.658	0.24
+ Micro cond.	20.02	0.498	0.673	0.41
+ RoPE	18.40	0.502	0.680	0.48
+ Internal VAE	19.57	0.492	0.668	0.52
+ Internal CLIP	**17.97**	0.607	0.717	0.65
+ Synth. captions	18.04	**0.685**	**0.751**	**0.81**

Table 2: Text-to-image model ablation studies.

Module	VBench		
	Quality ↑	Semantic ↑	Total ↑
Base model	80.19	70.51	78.25
w/ temp. patch=1	**80.92**	71.69	**79.07**
w/ temp. patch=4	79.72	69.15	77.61
w/ causal temp. attn	74.59	**73.13**	74.30
+ temp. mask	77.58	65.95	75.25
- spatial mask	80.57	70.31	78.52

Table 3: Ablation Study of T2V model design using T2V-XL. The base model uses temporal path size 2, non-causal temporal attention, spatial masking ratio 0.5, and no temporal masking.

Init.	MSRVTT	VBench		
	FVD ↓	Quality ↑	Semantic ↑	Total ↑
Scratch	417.98	80.27	67.84	77.78
T2V-256	415.63	80.28	71.29	78.49
T2I-512	**401.83**	79.77	71.58	78.13
Both	405.14	**80.45**	**72.37**	**78.83**

Table 4: Different model initialization for T2V-XL-512.

A.2   ABLATION STUDIES FOR KEY CHANGES ON T2I

We conduct a comprehensive ablation study to understand the impact of various model architecture designs and training strategies mentioned in Sec. 2 on the text-to-image generation task. To evaluate generation quality, we use a suite of popular automated metrics, including FID score [30], GenEval [24], DSGEval [13], and Image Reward [63].

We began with a base T2I-XL model, a DiT [44] model augmented with cross-attention layers to integrate with text embeddings. Initially, we applied a series of stabilization techniques, including QK-norm, sandwich-norm and singleton condition norm, which yielded comparable results to the baseline. Notably, these techniques enabled us to train models stably even with a learning rate increased from 1e-4 to 2e-4. We demonstrated that incorporating Flow Matching during training and employing CFG-Renormalization[2] during inference improved all the metrics substantially. Subsequently, we explored techniques to reduce training memory, such as AdaFactor Optimizer, MaskDiT, and Shared AdaLN, which maintained similar performance. Utilizing micro conditions and RoPE further reduced the FID score and improved DSGEval and Image Reward. Finally, incorporating an internally trained bigG CLIP model improved on all metrics even more. Notably, combining synthetic recaption with original caption following [35] achieved the best results in almost all metrics. For more details, refer to the Appendix G.

We use the optimal model architecture and training hyperparameters based on the T2I ablation study as our starting point for the remaining T2V and TI2V experiments.

A.3   ABLATION STUDIES ON KEY DESIGNS FOR T2V

Key Modulation We make some design choices in our model based on the evaluations on VBench, as shown in Tab. 3. The base model uses a temporal path size of 2, non-causal temporal attention, and a spatial masking ratio of 0.5. As expected, the model with temporal patch=1 is slightly better albeit with double compute. Using temporal patch=4 leads to a noticeable performance drop. Using causal temporal attention also results in a significant drop in both quality and total scores. Furthermore, removing the spatial masking results in a slight decrease in the Semantic score and an improvement in the Quality and Total scores. However, this comes at the cost of requiring more compute as the length of tokens are doubled. On the other hand, using temporal masking significantly degrades model performance, with large drops observed in the VBench quality and final scores.

Model Initialization We investigate how initialization impacts the performance of T2V-XL models. We train 512^2 T2V models by four different paths under a controlled total FLOPs setting: from

---
[2]Detailed description in Appendix.

scratch, initializing from a lower resolution T2V-256 model, initializing from a T2I-512 model, and loading both the temporal and spatial weights from T2V-256 and T2I-512 models respectively (Tab. 4). We find that jointly initializing from both a low resolution T2V model and a high resolution T2I model can achieve better VBench metrics. This joint initialization method yields slightly improved FVD values compared to training from scratch and offers benefits in terms of efficient experimentation and cost when low resolution models are already present.

A.4 ABLATION STUDIES ON KEY DESIGNS FOR TI2V

Here, we studied each design component in TI2V framework and tackled multi-task learning and staleness issue encountered when training high resolution TI2V models.

A.4.1 THE EFFECT OF FRAME REPLACEMENT

Models	Subj Cons	Bg Cons	Temp Flick	Mot Smooth	Dyn Deg	Aesth Qual	Img Qual	I2V Subj	I2V Bg	Cam Mot	I2V Avg Scores	Avg Scores
CA	82.2	92.8	95.7	96.3	**42.4**	48.8	65.5	88.9	90.9	26.9	68.2	73.0
CA + LP	95.2	**98.7**	**97.4**	98.1	22.2	57.3	**66.8**	96.9	97.3	22.7	72.3	75.3
FR	94.5	98.3	96.6	97.8	36.6	**58.0**	66.1	96.8	97.1	**31.5**	**75.8**	**77.3**
FR + CA	95.1	98.6	97.0	98.1	35.4	**58.0**	66.2	96.9	97.3	28.8	74.4	77.1
FR + CA + LP	**95.3**	98.5	97.3	**98.2**	22.4	57.3	66.3	**97.0**	**97.4**	25.8	73.4	75.6

Table 5: Ablation Study Results for Different Model Components for Text-Image-To-Video (TI2V) task on VBench-I2V.

We ablate several model variants for TI2V on STIV-XL model, by combining the following key components: Frame Replacement (FR), Cross Attention (CA), and Large Projection (LP). As shown in Tab. 5, notably, adding a large projection layer enhances the information passed by the cross-attention, resulting in improvements in both subject and background consistency. However, this approach may overly constrain the model, as evidenced by a reduction in the dynamic degree score (22.36 for FR + CA + LP compared to 35.4 for FR + CA), indicating that the model might exert excessive control over the generated output. In contrast, frame replacement alone has proven to be a robust and effective approach, yielding consistent improvements without negatively impacting other dimensions of video quality. The frame replacement (FR) model achieves high scores in I2V average scores (the average of I2V Subject, I2V Background and Camera Motion) and total average scores. These results underline the advantage of frame replacement as a foundational component, providing a stable backbone for maintaining quality across diverse dimensions.

A.4.2 THE EFFECT OF IMAGE CONDITION DROPOUT

Our experiments show that image condition dropout with JIT-CFG not only supports multi-task training but also resolves staleness in a 512^2 STIV model.

Multi-task training By using image-conditioning dropout during STIV training, we effectively enable both T2V and TI2V capability. As shown in Tab. 6, models trained exclusively on T2V or TI2V task alone cannot perform the other task, while STIV with image condition dropout can easily handles both two task well, achieving performance comparable to the best single-task models.

Model (M-512)	VBench-T2V			VBench-I2V		
	Q ↑	S ↑	T ↑	I ↑	Q ↑	T ↑
T2V	82.2	77.0	**81.2**	/	/	/
STIV	74.6	31.9	66.1	**98.0**	82.1	**90.1**
STIV-JIT	82.3	**74.1**	80.7	97.6	81.9	89.8
STIV-JIT-TUP	**83.0**	73.1	81.0	97.2	**82.3**	89.7

Table 6: Comparison of T2V, STIV and STIV with JIT-CFG on VBench and VBench-I2V **I**2V Score, **Q**uality, **T**otal scores.

Greater motion In practice, we have observed that while STIV-M-512 performs well on VBench-I2V, it sometimes generates static frames. The VBench-I2V metric tends to favor videos with less

motion, prioritizing smoothness and consistency. As shown in Tab. 7, STIV with JIT-CFG achieves higher dynamic degrees at the cost of a slight reduction in consistency and smoothness scores. We also show visual comparisons from Fig. 7 to Fig. 9 in the Appendix.

Model	Dyn Deg	Mot Smooth	Temp Flick	Bg Cons
STIV-M-512	10.2	**99.6**	**99.3**	**99.1**
STIV-M-512-JIT	**24.0**	99.1	98.6	98.6

Table 7: Effect of JIT-CFG on motion-related scores.

B RESULTS

Based on all of these studies, we scale our T2V and STIV model from 600M to 8.7B. We show the main results in Table 8 and Table 9, comparing our models with state-of-the-art open sourced and close sourced models, which demonstrates the effectiveness of our recipes. Specifically, we do finetuning on top of the pretrained video generation models (SFT), based on the 20,000 videos filtered from Panda-70M [11] using the method mentioned in Section G.1. Since we adopt MaskDiT technique in our pretraining stage, we try finetuning our model in an unmask manner (UnmaskSFT). We also finetuned our STIV model to become a temporal upsampler to interpolate the videos generated by our main T2V and STIV models to boost the motion smoothness (+ TUP).

T2V Performance We first showcase the effectiveness of our T2V model as the foundation for STIV. Table 8 presents a comparison of different T2V model variants on VBench, including the VBench-Quality, VBench-Semantic, and VBench-Total scores. Our analysis reveals that scaling up model parameters in our T2V model improves semantic following capability. Specifically, as model size increase from XL to XXL and M, VBench-Semantic scores rise from 72.5 to 72.7 and then to 74.8. This explicit emergence (from XL, XXL to M), suggesting larger models are better at capturing semantic information. However, the impact on video quality, measured by VBench-Quality, remains modest, with only a slight increase from 80.7 to 81.2 and then to 82.1. This finding suggests that scaling has a greater effect on the model's semantic capabilities than on video quality. Furthermore, increasing the spatial resolution from 256 to 512 significantly boosts the VBench-Semantic score from 74.8 to 77.0. Detailed results can be found in Table 12.

The Influence of SFT Additionally, fine-tuning the model with high-quality SFT data markedly enhances the VBench-Quality score from 82.2 to 83.9. Finetuning our model without any masking slightly increases the performance of model on the semantic score. Our best model achieves a VBench-Semantic score of 79.5, outperforming closed source models such as KLING, PIKA, and Gen-3. With the temporal upsampler, our model can achieve the state-of-the-art quality score compared with all other models.

TI2V Performance As shown in Tab. 9, our model delivers competitive performance compared to state-of-the-art approaches. It also reveals that while scaling up improves the I2V score, it has minimal impact on quality. In contrast, increasing the resolution leads to noticeable improvements in both quality and I2V scores. We provide complete results for the decomposed dimensions in Table 13.

C RELATED WORK

Text-To-Video Generation In recent years, diffusion-based methods have emerged as the dominant approach in text-to-video generation, both for close-source models [43, 45, 47] and open-source models [66, 72]. [28, 8, 9] leverages latent diffusion models (LDMs) [51] to enhance training efficiency. VideoLDM [5] integrates temporal convolution and attention mechanisms into the LDM U-Net for video generation. Recently, there has been a shift from UNet to diffusion transformer-based architectures [47, 66, 73, 22]. CogVideoX [65] adopts the framework from SD3 [20] to incorporate self-attention on the entire 3D video sequence with text conditions. Lumina-T2X [39] employs zero-init attention to transform noise into different modalities. In contrast to previous models, our focus is

Model	Quality ↑	Semantic ↑	Total ↑
Open Sourced Models			
OpenSora V1.2	81.4	73.4	79.8
AnimateDiff-V2	82.9	69.8	80.3
VideoCrafter-2.0	82.2	73.4	80.4
T2V-Turbo	82.2	74.5	80.6
CogVideoX-2B	82.2	75.8	80.9
Allegro	83.1	73.0	81.1
CogVideoX-5B	82.8	77.0	81.6
LaVie-2	83.2	75.7	81.8
Close Sourced Models			
Gen-2	82.5	73.0	80.6
PIKA	82.9	71.8	80.7
EMU3	84.1	68.4	81.0
KLING	83.4	75.7	81.9
Gen-3	84.1	75.2	82.3
Ours			
XL	80.7	72.5	79.1
XXL	81.2	72.7	79.5
M	82.1	74.8	80.6
M-512	82.2	77.0	81.2
+ SFT	83.9	78.3	82.8
+ SFT + TUP	84.2	78.5	**83.1**
+ UnmaskSFT	83.7	**79.5**	82.9
+ UnmaskSFT + TUP	**84.4**	77.2	83.0

Table 8: Performance comparison of T2V variants with open-sourced and close-sourced models on VBench.

Model	Quality ↑	I2V ↑	Total ↑
VideoCrafter-I2V	81.3	89.0	85.1
Consistent-I2V	78.9	94.8	86.8
DynamicCrafter-256	80.2	96.6	88.4
SEINE-512	80.6	96.3	88.4
I2VGen-XL	81.2	95.8	88.5
DynamicCrafter-512	81.6	96.6	89.1
Animate-Anything	81.2	**98.3**	89.8
SVD	**82.8**	96.9	89.9
STIV-XL	79.1	95.7	87.4
STIV-M	78.8	96.3	87.6
STIV-M-512	<u>82.1</u>	<u>98.0</u>	**90.1**
STIV-M-512-JIT	81.9	97.6	89.8

Table 9: Performance comparison of STIV-TI2V variants with open-sourced and close-sourced models on VBench-I2V.

to scale our diffusion transformer-based video generation model with spatial, temporal, and cross attention to over 8B parameters using various techniques. This model achieves good performance on VBench and serves as a strong baseline for the development of our text-image-to-video model: STIV.

Text-Image-To-Video Generation Controlling video content solely through text poses significant challenges in achieving satisfactory alignment between the video and the input text, as well as fine-grained control over the video generation process. To address this issue, recent approaches have integrated both the first frame and text to enhance control over video generation [70, 25, 62, 8, 50], mostly based on U-Net architecture. I2VGen-XL [70] builds upon the SDXL and employs a cascading technique to generate high-resolution video. DynamiCrafter [62] and VideoCrafter [8] use cross-attention to incorporate image condition. ConsistentI2V [50] employs a similar frame replacement strategy, but it also requires spatial temporal attention over the initial frame and special noise initialization to enhance consistency. Animate Anything [16] also employs the frame replacement technique, but it requires the use of motion strength loss to enhance the motion. However, their

Dynamic Degree on VBench-I2V is relatively low, at 2.7%. We apply frame replacement on the DiT architecture with our proposed image condition dropout method, and JIT-CFG can generate high quality I2V videos while effectively addresses the motion staleness issue.

D   JOINT IMAGE-TEXT CLASSIFIER-FREE GUIDANCE

We introduce a novel framework, Joint Image-Text Classifier-Free Guidance (JIT-CFG), in Section A.4, which facilitates the seamless integration of text and image conditions to enhance the modeling performance. This is accomplished through a modified velocity estimate, expressed as:

$$\hat{\boldsymbol{F}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t) = \boldsymbol{F}_\theta(\boldsymbol{x}_t, \varnothing, \varnothing, t)$$
$$+ \omega \cdot (\boldsymbol{F}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t) - \boldsymbol{F}_\theta(\boldsymbol{x}_t, \varnothing, \varnothing, t))$$

The approach employs text and image condition dropout, which is also critical for unifying T2V and TI2V tasks.

Probability mass shift Our model learns $P(\boldsymbol{x}|\boldsymbol{c}_T, \boldsymbol{c}_I)$, the probability distribution of generating video $\boldsymbol{x}$ given the text prompt $\boldsymbol{c}_T$ and image condition $\boldsymbol{c}_I$. Here, we demonstrate how JIT-CFG shifts the probability mass toward regions of higher likelihood, conditioned on $\boldsymbol{c}_T$ and $\boldsymbol{c}_I$. First, consider a score-matching model with JIT-CFG

$$\hat{\boldsymbol{s}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t) = \boldsymbol{s}_\theta(\boldsymbol{x}_t, t)$$
$$+ \omega \cdot (\boldsymbol{s}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t) - \boldsymbol{s}_\theta(\boldsymbol{x}_t, t))$$

Using the definition of score and *Bayes' Rule*, we derive

$$\hat{\boldsymbol{s}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t)$$
$$= \nabla log P_t(\boldsymbol{x}_t) + \omega \cdot (\nabla log P_t(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I) - \nabla log P_t(\boldsymbol{x}_t))$$
$$= \nabla log P_t(\boldsymbol{x}_t) + \omega \cdot \nabla log P(\boldsymbol{c}_T, \boldsymbol{c}_I|\boldsymbol{x}_t)$$
$$= \nabla log (P_t(\boldsymbol{x}_t) P_t^\omega(\boldsymbol{c}_T, \boldsymbol{c}_I|\boldsymbol{x}_t))$$
$$= \nabla log (P_t^{1-\omega}(\boldsymbol{x}_t) P_t^\omega(\boldsymbol{x}_t|\boldsymbol{c}_T, \boldsymbol{c}_I)),$$

where ω determines the influence of the text and image conditions during sampling from the tempered distribution. For a flow-matching model employing linear interpolants, the velocity and score are related as [42]:

$$\hat{\boldsymbol{s}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t) = \frac{t}{1-t}\hat{\boldsymbol{F}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t) - \frac{1}{1-t}\boldsymbol{x}_t,$$

It implies $\frac{\partial \hat{\boldsymbol{s}}}{\partial \hat{\boldsymbol{F}}} = \frac{t}{1-t} > 0$, meaning that the JIT-CFG-guided velocity $\hat{\boldsymbol{F}}$ shifts the probability mass in alignment with the modified score $\hat{\boldsymbol{s}}$ by adjusting the tempered distribution.

CFG-Renormalization Empirically, we observed that the magnitude of the modified velocity, $||\hat{\boldsymbol{F}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t)||$ tends to be very large during the early stages of integration in inference (i.e. when t is small). This behavior sometimes leads to overshooting beyond the learned latent distribution, resulting in artifacts in the generated output. We identified this issue as primarily due to the significant difference between the conditional velocity, $\boldsymbol{F}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t)$, and the unconditional velocity, $\boldsymbol{F}_\theta(\boldsymbol{x}_t, \varnothing, \varnothing, t)$ when t is small.

To mitigate this, we propose a simple yet effective renormalization method that re-scales the magnitude of the modified velocity to $||\boldsymbol{F}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t)||$ while preserving its direction. Formally, this is defined as:

$$\tilde{\boldsymbol{F}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t) = ||\boldsymbol{F}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t)|| \frac{\hat{\boldsymbol{F}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t)}{||\hat{\boldsymbol{F}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t)||}$$

As shown in Table 2, this technique significantly improves performance across various T2I evaluation benchmarks.

E   Implementation Details for T2V and STIV

Given that we use spatial-temporal attention, we first pretrain 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 JIT-CFG setting, we also independently randomly drop image condition 8% during training. For both T2V and STIV models, the JIT-CFG scale is set to 7.5. The training schedule follows the progressive training recipe described in section 2.2.

F   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.

G   Ablation Study on Text-to-Image Generation

Baseline Setup For our base model, we employed the PixArt-α architecture [10], which builds on the DiT [44] 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[3] and OpenAI CLIP L14 model[4], 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 2 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 [29] to manage the activation scale within attention layers. Additionally, we applied sandwich-norm [26] 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 resulting in quality improvements.

Noising and Denoising Process Formulation We explored optimized noising/denoising formulations by replacing the diffusion process with a flow-based linear interpolant[42]. 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. D.

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 2, masked training may adversely affect metrics such as FID and HPS. However, we found additional unmask finetuning for a

---

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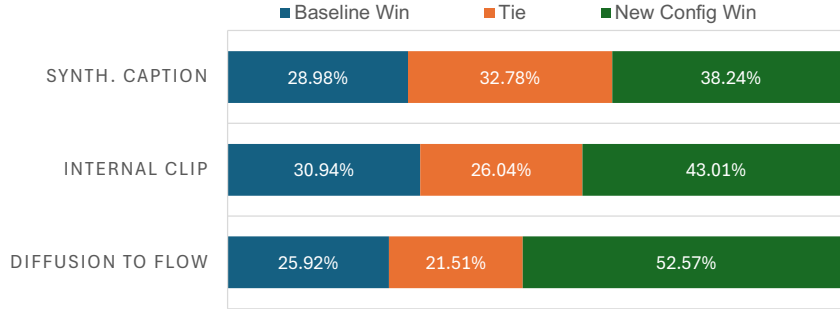[4]https://huggingface.co/openai/clip-vit-large-patch14

Figure 4: Human evaluation results on significant changes in T2I ablation study Tab. 2.

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 [36] 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 [46], to mitigate cropping artifacts in elongated objects. Finally, synthetic captions generated via [35] were included in our training data, resulting 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 our usage of automatic evaluation as daily development metrics to maintain generation quality and prevent regressions leading to significant quality losses.

G.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 [11] as a working example and produce a curated subset, Panda-30M, via our pipeline.

Video Pre-processing and Feature Extraction. We employ PySceneDetect[5] 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

---

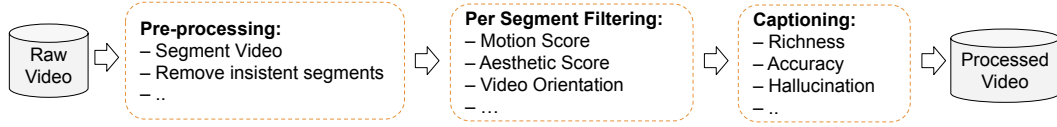[5]https://github.com/Breakthrough/PySceneDetect

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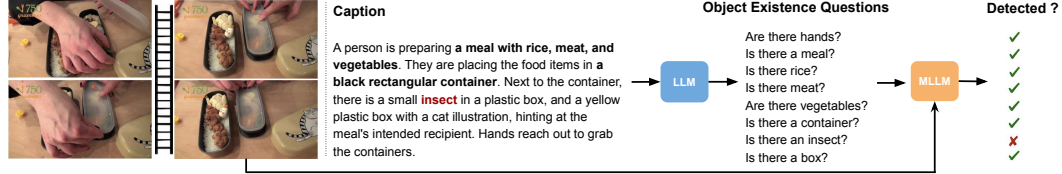

Figure 6: An overview of DSG-Video's approach to detecting object hallucinations in captions: we use an LLM to generate questions and another MLLM to validate the presence of the object across frames. If the MLLM fails to detect the object in all frames, the object is classified as a hallucination.

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. 10: instead of pursuing data volume, higher-quality videos have the potential to achieve more promising results.

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

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 [4], 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 [69]—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 [13], 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 demonstrate that richer, more accurate video descriptions can significantly enhance downstream generation quality.

# H   DETAILED RESULTS FOR IMAGING DROPOUT

As mentioned in Section A.4.2, after adding imaging dropout. We observe this phenomenon happens when we scale our model to 8B with >= 512 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 [47] As shown in Fig. 7 to 9, using image condition dropout in general achieves better performance than the baseline in terms of motion quality.

STIV-M-512 V.S. STIV-M-512-JIT

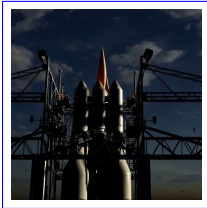**Prompt:** *A red panda taking a bite of a pizza.*
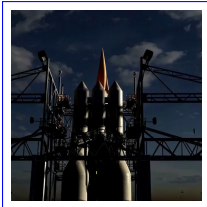

Reference Image
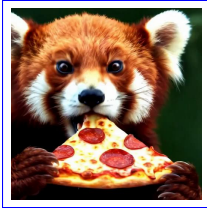
Reference Image

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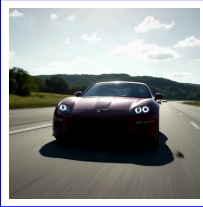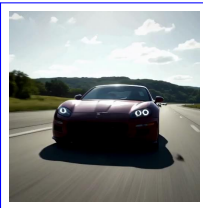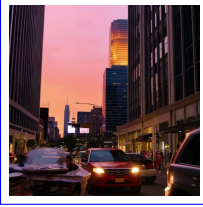

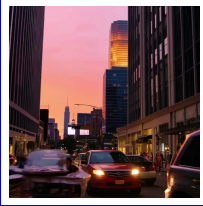**Reference Image**

Reference Image

Figure 7: Visualization of STIV-M-512 V.S. STIV-M-512-JIT. (Given the same prompt, the figures in the top row are generated by STIV-M-512, while the figures in the bottom row are generated by STIV-M-512-JIT.)

STIV-M-512 V.S. STIV-M-512-JIT

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.*
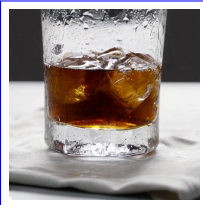
Reference Image
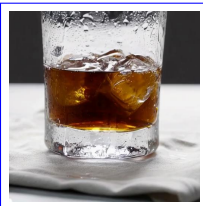
Reference Image

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

STIV-M-512 V.S. STIV-M-512-JIT

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.*
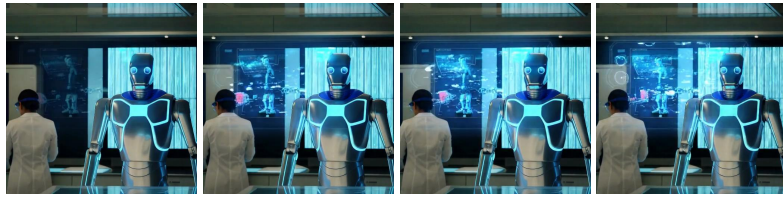

Reference Image

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.*

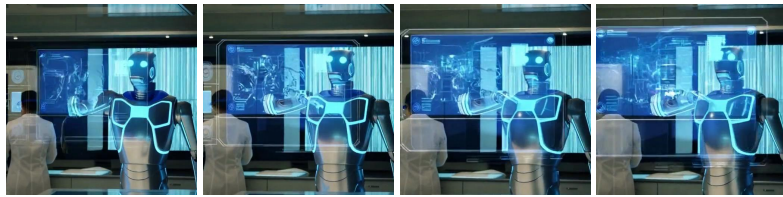**Reference Image**

Reference Image

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

Caption	Total Object	DSG-Video$_i$($\downarrow$)	DSG-Video$_s$($\downarrow$)	MSRVTT FVD ($\downarrow$)	VBench ($\uparrow$)
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.

I   DETAILED RESULTS FOR T2V AND STIV

I.1   DETAILS OF VBENCH AND VBENCH-I2V EVALUATION METRICS

We follow the same as the evaluation protocol provided by VBench [32].

I.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.

I.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.

I.2   DETAILED RESULTS ON VBENCH AND VBENCH-I2V

We showcase the detailed version of the performance shown in Tab. 12 and Tab. 13.

Model	Subject Cons.	Back. Cons.	Temporal Flickering	Motion Smooth.	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class	Multiple Objects	Human Action
CogVideoX-5B [65]	96.2	96.5	98.7	96.9	80.0	62.0	62.9	85.2	62.1	99.4
CogVideoX-2B [65]	96.8	96.6	98.9	97.7	59.9	60.8	61.7	83.4	62.6	98.0
Allegro [73]	96.3	96.7	99.0	98.8	55.0	63.7	63.6	87.5	59.9	91.4
AnimateDiff-V2 [27]	95.3	97.7	98.8	97.8	40.8	67.2	70.1	90.9	36.9	92.6
OpenSora V1.2 [72]	96.8	97.6	**99.5**	98.5	42.4	56.9	63.3	82.2	51.8	91.2
T2V-Turbo [38]	96.3	97.0	97.5	97.3	49.2	63.0	**72.5**	94.0	54.7	95.2
VideoCrafter-2.0 [9]	96.9	98.2	98.4	97.7	42.5	63.1	67.2	92.6	40.7	95.0
LaVie-2 [61]	97.9	98.5	98.8	98.4	31.1	**67.6**	70.4	**97.5**	64.9	96.4
LaVIE [61]	91.4	97.5	98.3	96.4	49.7	54.9	61.9	91.8	33.3	96.8
ModelScope [60]	89.9	95.3	98.3	95.8	66.4	52.1	58.6	82.2	39.0	92.4
VideoCrafter [8]	86.2	92.9	97.6	91.8	**89.7**	44.4	57.2	87.3	25.9	93.0
CogVideo [31]	92.2	95.4	97.6	96.5	42.2	38.2	41.0	73.4	18.1	78.2
PIKA [45]	96.9	97.4	**99.7**	**99.5**	47.5	62.4	61.9	88.7	43.1	86.2
Gen-3 [53]	97.1	96.6	98.6	99.2	60.1	63.3	66.8	87.8	53.6	96.4
Gen-2 [52]	97.6	97.6	99.6	**99.6**	18.9	67.0	67.4	90.9	55.5	89.2
KLING [34]	**98.3**	97.6	99.3	99.4	46.9	61.2	65.6	87.2	68.1	93.4
EMU3 [25]	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 [65]	82.8	66.4	53.2	24.9	25.4	27.6	82.8	77.0	81.6	70.0
CogVideoX-2B [65]	79.4	69.9	51.1	24.8	24.4	26.7	82.2	75.8	80.9	68.3
Allegro [73]	82.8	67.2	46.7	20.5	24.4	26.4	83.1	73.0	81.1	67.5
AnimateDiff-V2 [27]	87.5	34.6	50.2	22.4	26.0	27.0	82.9	69.8	80.3	64.7
OpenSora V1.2 [72]	90.1	68.6	42.4	24.0	24.5	26.9	81.4	73.4	79.8	66.0
T2V-Turbo [38]	89.9	38.7	55.6	24.4	25.5	28.2	82.6	74.8	81.0	67.4
VideoCrafter-2.0 [9]	92.9	35.9	55.3	**25.1**	25.8	28.2	82.2	73.4	80.4	66.0
LaVie-2 [61]	91.7	38.7	49.6	**25.1**	25.2	27.4	83.2	75.8	81.8	67.6
LaVIE [61]	86.4	34.1	52.7	23.6	25.9	26.4	78.8	70.3	77.1	63.8
ModelScope [60]	81.7	33.7	39.3	23.4	25.4	25.7	78.1	66.5	75.8	62.4
VideoCrafter [8]	78.8	36.7	43.4	21.6	25.4	25.2	81.6	72.2	79.7	62.3
CogVideo [31]	79.6	18.2	28.2	22.0	7.8	7.7	72.1	46.8	67.0	52.3
PIKA [45]	90.6	61.0	49.8	22.3	24.2	25.9	82.9	71.8	80.7	66.1
Gen-3 [53]	80.9	65.1	54.6	24.3	24.7	26.7	84.1	75.2	82.3	68.5
Gen-2 [52]	89.5	66.9	48.9	19.3	24.1	26.2	82.5	73.0	80.6	66.1
KLING [34]	89.9	**73.0**	50.9	19.6	24.2	26.4	83.4	75.7	81.9	68.8
EMU3 [25]	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 12: Detailed Evaluation Results for Text-To-Video Generation Models.

Model	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality
DynamicCrafter-256 [62]	94.7	98.3	98.1	97.8	40.6	58.7
DynamicCrafter-512 [62]	93.8	96.6	95.6	96.8	69.7	60.9
Animate-Anything [15]	98.9	98.2	98.1	98.6	2.7	67.1
SVD [4]	95.5	96.6	98.1	98.1	52.4	60.2
SEINE-512 [12]	95.3	97.1	97.3	97.1	27.1	64.6
VideoCrafter-I2V [9]	97.9	98.8	98.2	98.0	22.6	60.8
Consistent-I2V [50]	95.3	98.3	97.6	97.4	18.6	59.0
I2VGen-XL [70]	94.2	97.1	98.3	26.1	26.1	64.8
STIV-M	95.4	98.9	97.2	98.1	32.1	59.0
STIV-M-512	99.5	99.3	99.5	99.6	10.2	62.5
STIV-M-512-JIT	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 [62]	62.3	97.1	97.6	20.9	80.2	88.4
DynamicCrafter-512 [62]	68.6	97.2	97.4	32.0	81.6	89.1
Animate-Anything [15]	72.1	98.8	98.6	13.1	81.2	89.8
SVD [4]	69.8	98.8	98.6	62.3	82.8	89.9
SEINE-512 [12]	71.4	97.2	96.9	21.0	80.6	88.4
VideoCrafter-I2V [9]	71.7	91.2	91.3	33.6	81.3	85.1
Consistent-I2V [50]	66.9	95.8	96.0	33.9	78.9	86.8
I2VGen-XL [70]	69.1	96.5	96.8	18.5	81.2	88.5
STIV-M	66.1	97.0	97.4	22.7	78.8	87.6
STIV-M-512	71.5	99.2	97.3	13.2	82.1	90.1
STIV-M-512-JIT	71.0	98.8	97.5	15.1	81.9	89.8

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

J 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 14 and 15). When controlling for FLOPs we take into account, the compute used to pretrain 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 16 and 17 show the VBench quality metrics for high resolution and high frame count XL sized models respectively.

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 14: 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 15: A breakdown of FLOPs for training high frame count T2V models. Unit: 10^{21}.

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 16: 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 17: Detailed VBench metrics of different model initialization methods for higher frame count T2V model training.

K 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. 18.

Model	Source	PC	SA	Avg.
OpenSora [72]	Open	35	21	28
SVD [4]	Open	34	37	35
CogVideoX-2B [66]	Open	39	40	39
LaVIE [61]	Open	36	45	41
VideoCrafter2 [9]	Open	36	47	41
CogVideoX-5B [66]	Open	41	57	49

Model	Source	PC	SA	Avg.
Gen-2 [52]	Closed	31	26	29
Pika [45]	Closed	33	25	29
Lumiere-T2V [2]	Closed	31	35	33
Lumiere-T2I2V [2]	Closed	25	46	35
Luma Dream Machine [40]	Closed	30	53	41.5
XL	Our	36	57	47
M-512	Our	**43**	**59**	**51**

Table 18: Performance of T2V models on VideoPhy [1].

L 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 [23] to adopt the Inception Score (IS) [54] and FVD for the evaluation[6].

Tab. 19 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.

Method	IS ↑	FVD ↓	VBench-Quality ↑
CogVideo [31]	50.5	626	-
TATS [23]	79.3	332	-
MMVG [21]	73.7	328	-
VideoFusion [41]	80.0	173	-
MAGVIT [67]	83.6	159	-
XL-128	93.4	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	106	70.3

Table 19: Performance of Class-to-Video Generation on UCF-101.

M 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.

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. 20, the video-to-video model (STIV-V2V) shows significantly lower FVD scores compared to the text-to-video model (T2V) on MSRVTT [64] test set and MovieGen Bench [47]. This result indicates that video-to-video models can achieve superior performance, which is promising for applications in 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. 21 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. 6, 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.

---

[6]Following our baselines (https://github.com/songweige/TATS/issues/13), we apply C3D [59] pre-trained on UCF-101 for the IS logits. For FVD, we adopt I3D [7] pre-trained on Kinetics-400 [33] to calculate the video embeddings.

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

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

Model	GSO [19] PSNR ↑	SSIM ↑	LPIPS ↓
Zero123++	21.200	0.723	**0.143**
STIV-TI2V-XL	**21.643**	**0.724**	0.156

Table 20: Comparison of T2V and V2V. Table 21: Performance of STIV-TUP. Table 22: Multiview generation comparison.

Here, we adopt the novel view camera definitions outlined in Zero123++ [56], 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 [17]. For a fair comparison, we increased the image resolution to 320 during finetuning, aligning with the settings used in Zero123++. Our evaluation used objects from the Google Scanned Objects dataset [19], where we compared the output multi-view images against ground-truth renderings. As shown in Fig. 22, 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 Figure 10.

Long Video Generation We develop an effective and efficient framework to generate long videos. Specifically, we propose a hierarchical framework, including training our STIV on two different modes: (1) key frame prediction by learning uniformly sampled video frames with stride of 20 with image conditioning dropout and (2) interpolated frame generation by learning consecutive video frames with first and last frame as image conditions. During sampling stage, we change the image and micro conditions, and first use the first mode to generate key frames and then generate the interpolated frames use the second mode, leading to a long video. It is natural to reuse the STIV model to autoregressively generate the videos conditioned on previous generated one. However, in practice, we found this rollout approach can be hurt by error propagation in the previous video, and lacks some global consistency between frames. Therefore, we propose a simple yet effective baseline, purely based on our STIV framework. As mentioned in the main text, we design a hierarchical framework, train our STIV on two different modes: (1) key frame prediction by learning uniformly sampled video frames with a stride of 20 with image conditioning dropout, and (2) interpolated frame generation by learning consecutive video frames with the first and last frame as image conditions. During the sampling stage, we change the image and micro conditions, and first use the first mode to generate key frames and then generate the interpolated frames using the second mode, leading to a long video. We showcase one long T2V and one TI2V example in Figure 11; we achieve $(20 - 1) \times 20 = 380$ frames in total. We uniformly sampled 8 frames out of the 380 frames. Noted that this is only an early exploration of long video generation, and we do not have many long enough videos in our training distribution, so we leave it as future work to further explore the architecture to boost long video synthesis.

N 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 [47] and Sora [43].
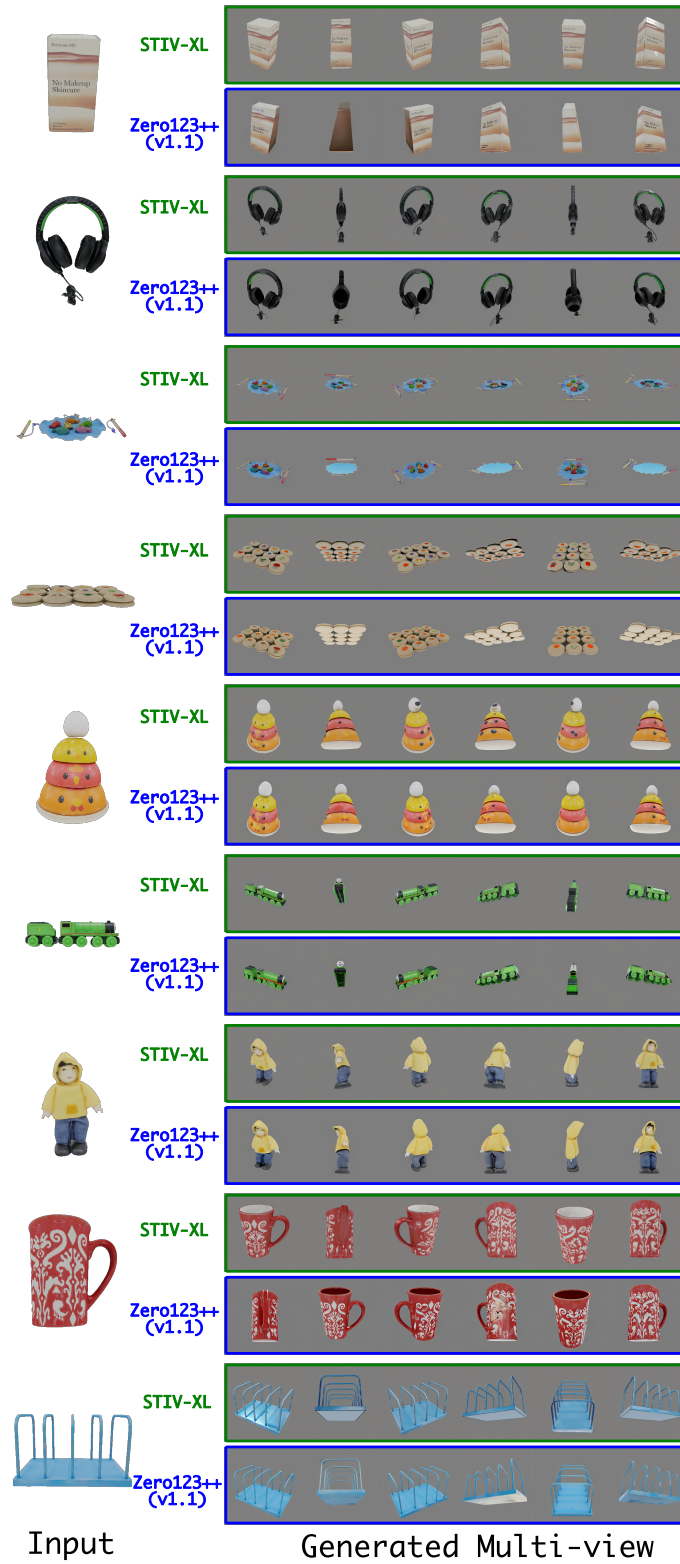
Figure 10: The visual comparison between our STIV-XL with Zero123++ [56] on GSO [19].

Long Video Generation

Prompt: *A drone camera circles around a beautiful historic church built on a rocky outcropping along the Amalfi Coast, the view showcases historic and magnificent architectural details and tiered pathways and patios...*

Reference Image

Prompt: *A slow cinematic push in on an ostrich standing in a 1980s kitchen.*

Figure 11: Visualizations of long video generation framework.

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.*
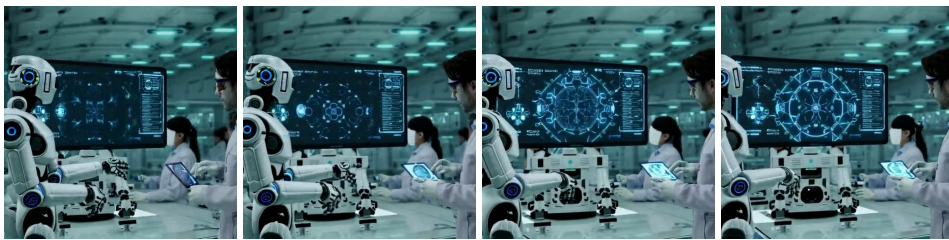
Prompt: *A robotic arm wielding a glowing sword, battling a shadowy figure in a high-tech dojo, each strike creating sparks that light up the space.*
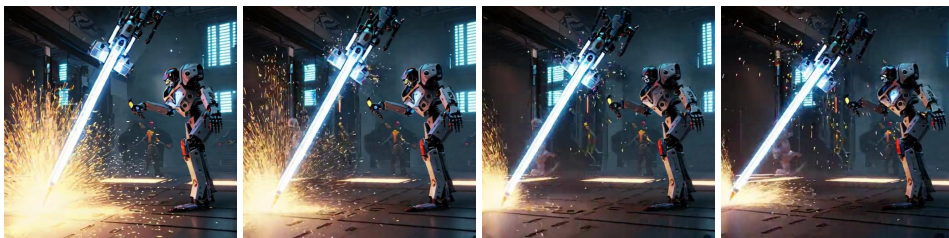
Prompt: *A city skyline reflected in the water, but the reflection shows an alternate world with flying cars, towering robots, and futuristic architecture.*

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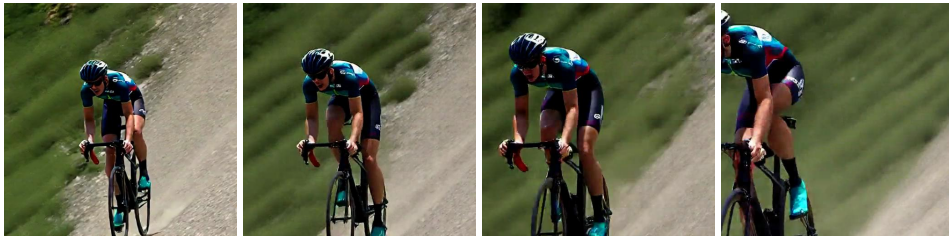
Prompt: *A speed skater accelerating during a short track race.*

Text-Image-to-Video

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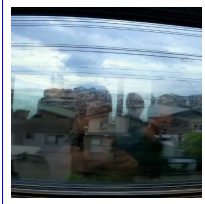
Reference Image

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...*
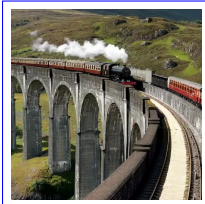
Reference Image

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...*
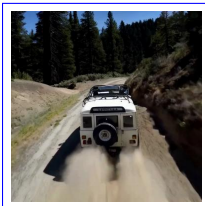
Reference Image

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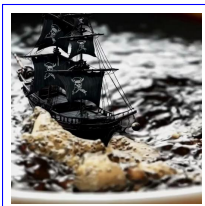
Reference Image

Text-Image-to-Video

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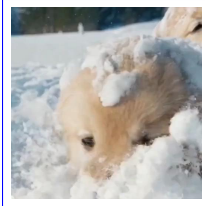
Reference Image

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.*
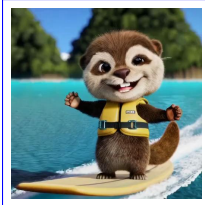
Reference Image

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.*
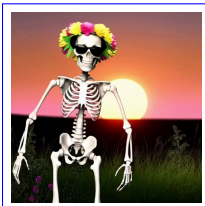
Reference Image

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...*
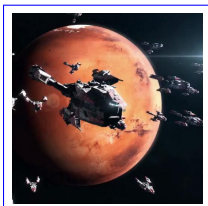
Reference Image

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...*
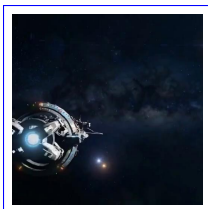
Reference Image
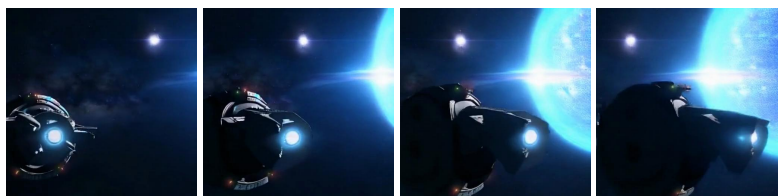
Prompt: *Robots move efficiently through a futuristic laboratory, adjusting holographic displays and conducting experiments, while scientists observe and interact with the high-tech equipment.*
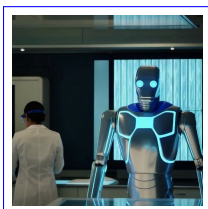
Reference Image
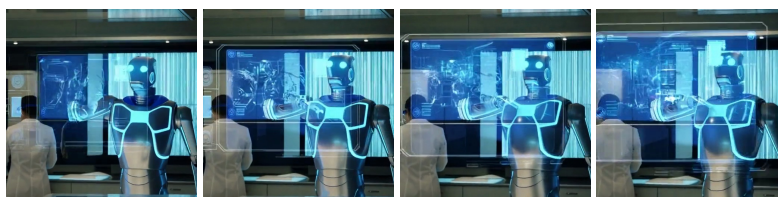
Prompt: *The video presents a serene scene with a group of camels walking in a line across a desert landscape. The camels are adorned with colorful saddles and are led by a person wearing a green garment. The background features a clear sky...*
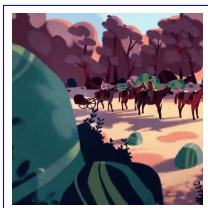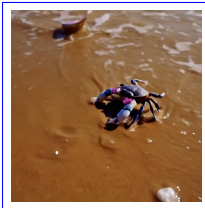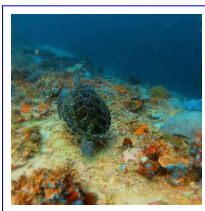

Reference Image
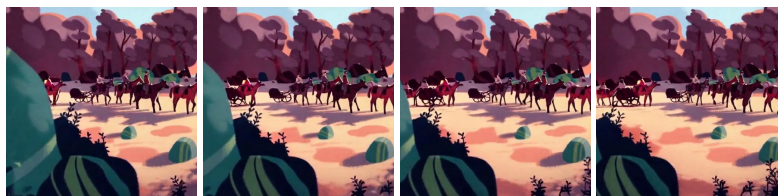
Text-Image-to-Video

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...*
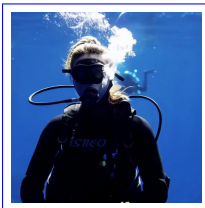

Reference Image

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...*
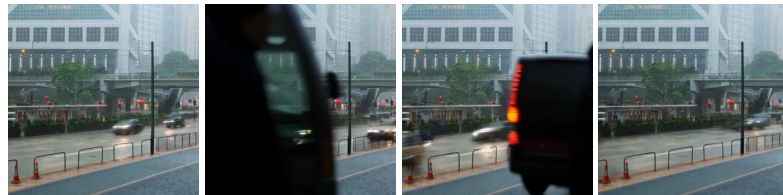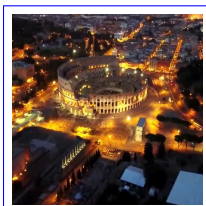

Reference Image

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...*
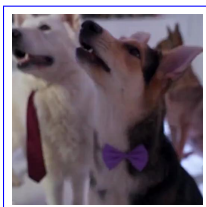

Reference Image

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