Inflation with Diffusion: Efficient Temporal Adaptation for Text-to-Video Super-Resolution

Xin Yuan[1]*, Jinoo Baek[2], Keyang Xu[2], Omer Tov[2], Hongliang Fei[2]
[1]University of Chicago [2]Google
yuanx@uchicago.edu {jinoo,keyangxu,omertov,hongliangfei}@google.com

Abstract

We propose an efficient diffusion-based text-to-video super-resolution (SR) tuning approach that leverages the readily learned capacity of pixel level image diffusion model to capture spatial information for video generation. To accomplish this goal, we design an efficient architecture by inflating the weightings of the text-to-image SR model into our video generation framework. Additionally, we incorporate a temporal adapter to ensure temporal coherence across video frames. We investigate different tuning approaches based on our inflated architecture and report trade-offs between computational costs and super-resolution quality. Empirical evaluation, both quantitative and qualitative, on the Shutterstock video dataset, demonstrates that our approach is able to perform text-to-video SR generation with good visual quality and temporal consistency. To evaluate temporal coherence, we also present visualizations in video format in google drive.

1. Introduction

Diffusion model [5, 14], as a deep generative model, has achieved a new state-of-the-art performance, surpassing GANs [3, 6, 16, 18] in generative tasks [9, 10]. In a diffusion-based text-conditioned generation system, a base model initially generates a low-resolution image/video, which is subsequently refined by a super-resolution module [4, 10, 12] to produce high-quality samples. Numerous existing diffusion-based text-to-image super-resolution models [9, 10], trained on billion-scale text-image dataset, have demonstrated outstanding generation capability. However, training text-to-video spatial super-resolution is challenging due to the scarcity of high-resolution video data. This scenario motivates the inflation of off-the-shelf image models to video generation tasks [2, 7, 15, 19]. Furthermore, training a video generation model needs exceedingly high computational and memory requirements, which drives techniques that offer cost-effective alternatives to optimize the video models.

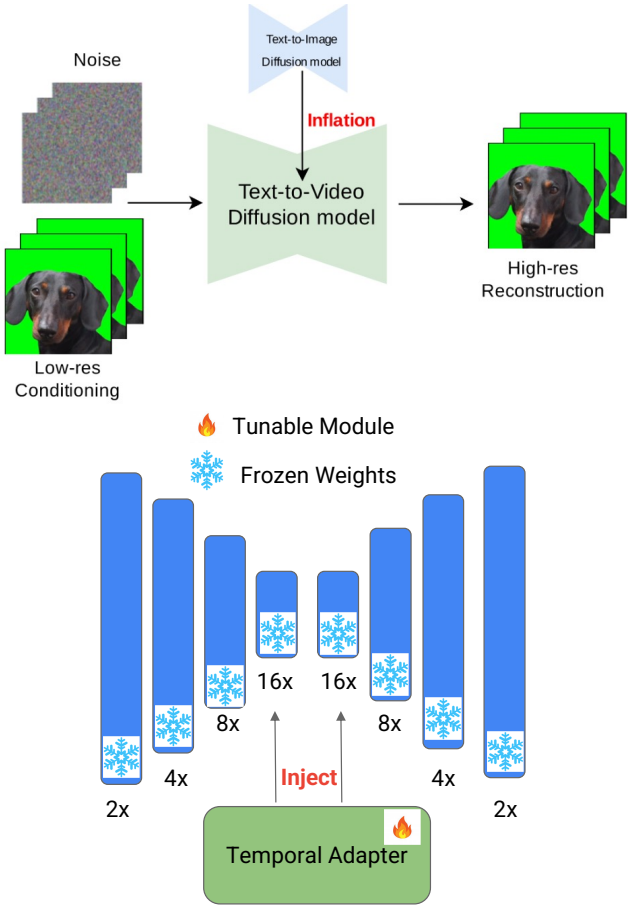


Figure 1: Overall architecture of our approach. *Up*: we inflate the UNet weights from a text-to-image model into a text-to-video model to perform a diffusion-based super-resolution task. *Bottom*: we inject and tune a temporal adapter in the inflated architecture while maintaining the UNet weights frozen.

Several recently proposed methods [1] also focus on generating high-quality videos using pretrained latent diffusion models. Temporal attention mechanisms are also commonly used in [4, 12]. Yet, investigating the trade-offs between video quality and the resource requirements in a fine-tuning stage is not the focus of those works. [1] typically requires

*This work has been done during the first author's internship at Google.

full tuning of all computational modules to generate high-quality videos, even with pretrained image weights inflated in the video architectures. In contrast, our approach lies in the applied domain and investigates how tuning efficiency affects the video super-resolution quality. More importantly, instead of investigating model inflation in latent space [1], our approach is the first to directly work on pixels. Note that our goal is not to achieve state-of-the-art generation quality. Instead, we aim to establish a practical and efficient tuning system to generate high-resolution videos with reasonable visual quality and temporal consistency.

In this paper, we aim to leverage the readily learned spatial capacity of image weights for efficient and effective text-to-video super-resolution, as shown in the upper of Figure 1. To capture the coherence across video frames, we inject an attention-based temporal adapter into the video architecture. This adapter can be fine-tuned independently while keeping inflated weights frozen, as shown in the bottom of Figure 1. We perform the spatial super-resolution task on the Shutterstock video dataset and validate that our approach is capable of generating videos with good visual quality and temporal consistency. We also demonstrate the trade-off between tuning complexity and generation quality.

2. Related Work

Diffusion-based SR model is conditioned on low-resolution samples, generated by a base generation model, to further produce high-resolution images [10] or videos [4]. With the success of image generation models pre-trained on billion-scale image data, recent research efforts have been made to directly borrow off-the-shelf image models for video tasks. For example, [7] load image weights for video generation in a zero-shot manner. [2, 15, 19] adopt model inflation and DDIM [13] inversion for text-to-video editing. While these studies may not directly apply to video spatial super-resolution task, they provide insightful hints on the feasibility of adopting an image model without necessitating re-training from scratch.

Temporal attention mechanisms that operate on the time axis are commonly adopted in video diffusion approaches [4, 12]. Our method shares the same concept with [1] in the spirit of borrowing image diffusion models for video generation. However, our approach focus on the applied domain for text-to-video super resolution. More importantly, with facilitating partial tuning of the video architectures, we qualitatively and quantitatively evaluate how different tuning methods affect the generation quality, including visual quality and temporal consistency.

3. Approach

Consider a video clip represented as a sequence of n image frames, denoted as $I : [I_1, ..., I_n]$, with low spatial resolution s, and a text description t for this clip, our objective is to generate a new video clip of the same length but with an enhanced resolution s' while preserving the correlation between text and video. We aim to exploit the robust spatial understanding of a pre-trained and fixed large-scale image diffusion model, repurposing it for video generation that remains temporally consistent. This removes the need for extensive training from scratch on limited high-resolution video data, which is both time- and resource-consuming. We achieve this goal by inflating the weights of image diffusion model into a video generation architecture (as detailed in Section 3.1) and further tuning an efficient temporal adapter to ensure the continuity and coherence across video frames, discussed in Section 3.2.

3.1. Inflation with Image Weights

We build a one-on-one mapping between image and video architectures through 'upgrade' Imagen [10] text-to-image super-resolution model to accommodate video tasks. We first revisit the U-Net architecture in Imagen, composed of residual and cross-attention blocks, as shown in Figure 2 (left). Given a batch of input static images with shape $B \times C \times H \times W$, the residual blocks capture the spatial information while cross-attention ensures that the generated image aligns with the given text prompt. In the context of our text-to-video super-resolution, the input batch of video clips is in the shape of $B \times F \times C \times H \times W$, where F is the number of frames. As shown in Figure 2 (right), each individual frame is processed through a parallel scheme, each branch contains a residual block and cross-attention layer for text-visual feature extraction. At the end of the UNet block, we have a temporal adapter for feature aggregation to maintain consistency and smoothness across frames. The processing units of the residual block and cross-attention layer share the same weights during training, in which case we can simply reshape the video data into $(BF) \times C \times H \times W$. Given this nice property of weight sharing scheme, we can directly inflate the pre-trained image model weights into the video UNet without any architectural modification.

3.2. Temporal Adaptater with Frame-wise Attention

To capture the coherence among video frames, we apply the temporal adapter after the residual and cross-attention blocks. Figure 3 depicts the design of the attention-based temporal adapter. We first reshape video data I into I' with the shape of $B \times F \times (CHW)$ and then adopt a conventional self attention module:

$$\text{Self-Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V. \quad (1)$$

Such an attention mechanism is effective in determining the overall structure and the coherence of the video frames.

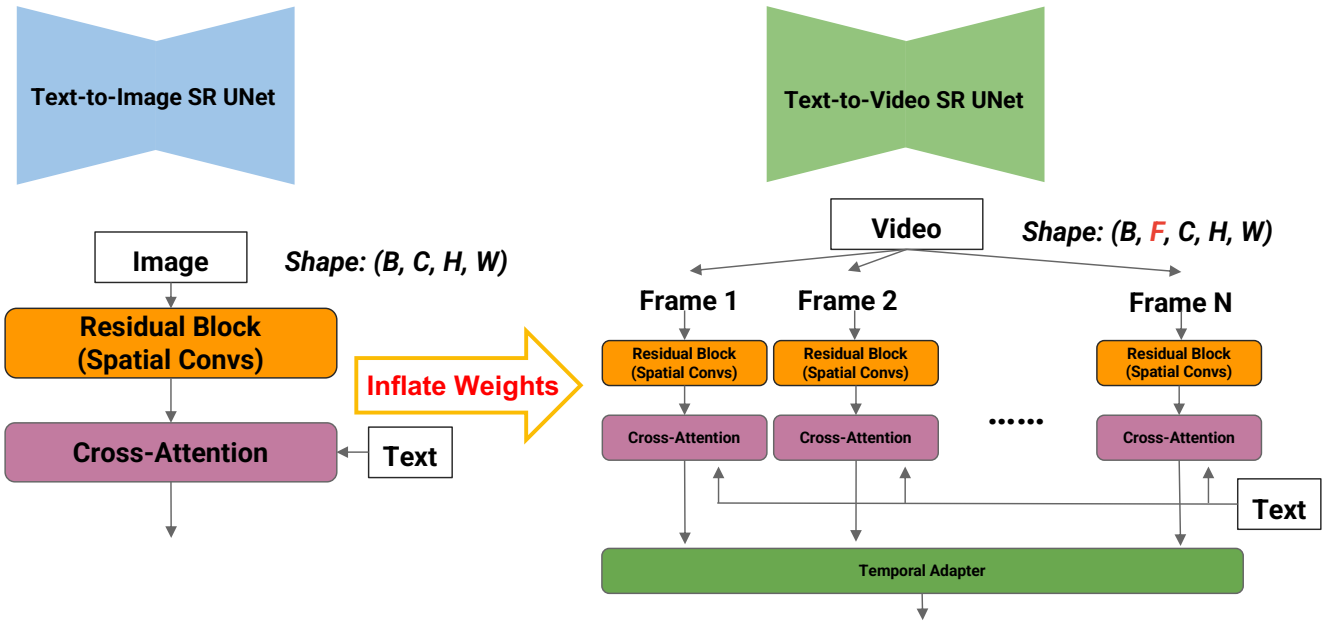


Figure 2: Weights inflation from a text-to-image SR UNet to a text-to-video SR UNet.

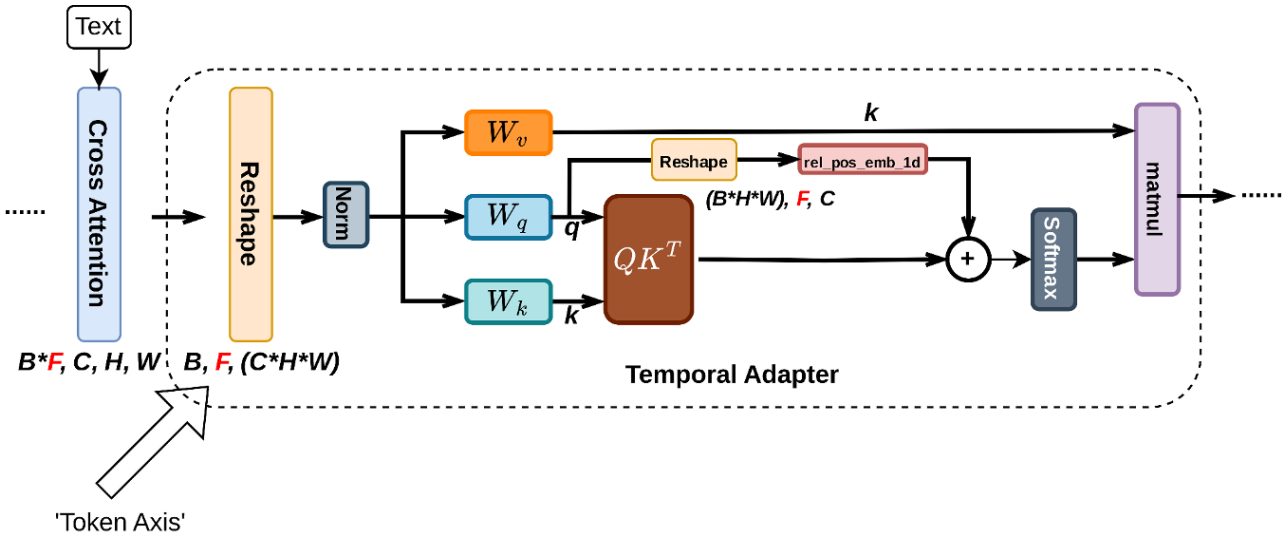


Figure 3: Temporal adapter with attention that ensures temporal coherence across a video clip.

Specifically, a weighted sum over the 'token axis' F is calculated to learn the frame-wise correlation. We employ end-to-end optimization of either the full or partial model weights, aligning with the simple denoising objective in DDPM [5]. As such, the model weights are optimized by minimizing the MSE loss of noise prediction, conditioned on the low-resolution frames.

4. Experiments

We validate our approach on the Shutterstock dataset. We inflate a version of the Imagen diffusion model pre-trained on our internal data sources for 8x super-resolution, into our video UNet. This UNet consists of four stages each for downsampling and upsampling, denoted as $2\times$, $4\times$, $8\times$, $16\times$. T5-xxl encoder [8] is used to extract text embedding, the output of which is fed into the cross-attention layers within the $16\times$ stage. We train our video model on the Shutterstock text-to-video dataset with 7 million video clips in a resolution of 256×256-resolution and frame rate of 8 FPS. The duration for each clip is 1 second, i.e. $F = 8$. The supper-resolution scale is $4\times$, elevating the resolution from 64×64 to 256×256. We investigate several baseline optimization approaches, including (1) Zero-shot (ZS):

Figure 4: Visualization of different tuning methods after image model inflation, conditioned on text prompt "Dog dachshund on chromakey".

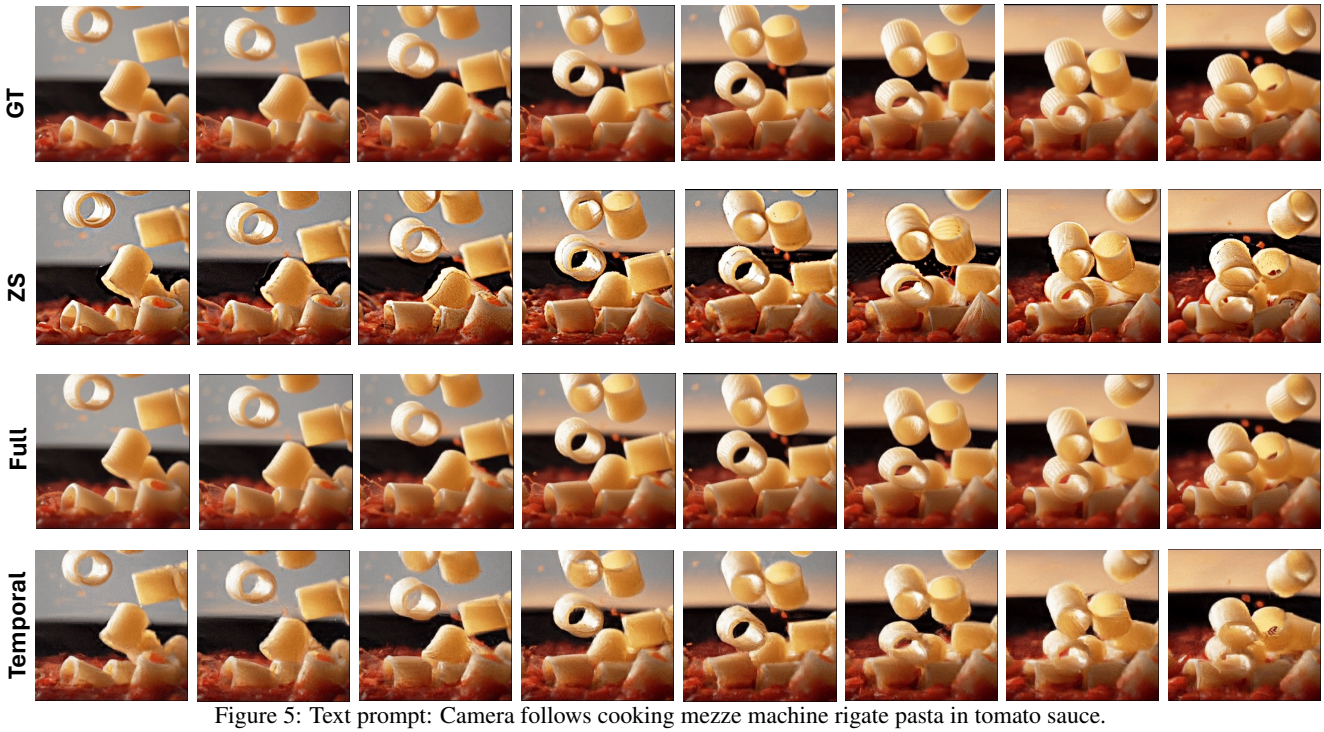


Figure 5: Text prompt: Camera follows cooking mezze machine rigate pasta in tomato sauce.

we directly evaluate the video model after inflation without further training. (2) Full-ft (Full): After integrating the temporal adapter, all modules undergo optimization. This strategy aims to showcase the potential 'upper bound' performance in the super-resolution task. (3) Temporal: we only tune the temporal adapter to capture the temporal consistency while maintaining superior generation quality efficiently. We finetune for 1 epoch, using Adafactor [11] with initial LR of 10^{-5} and batch size of 256 on 64 TPUv3.

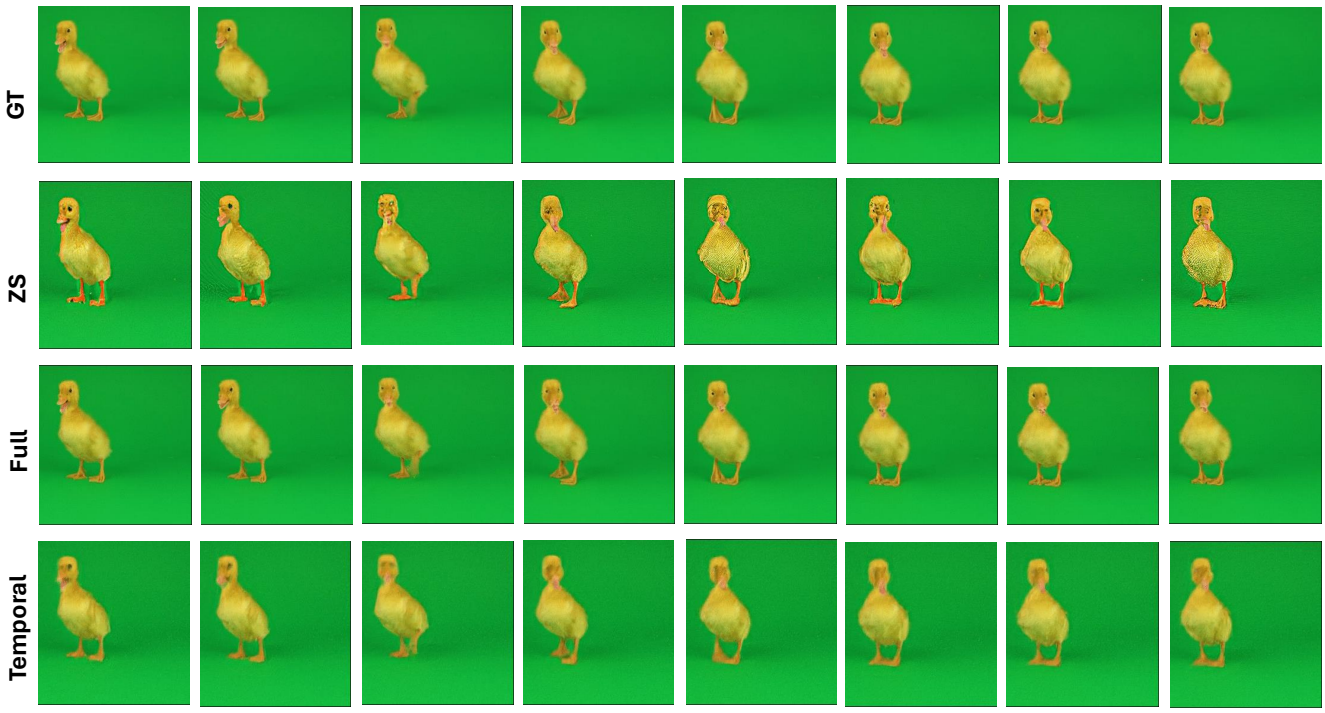


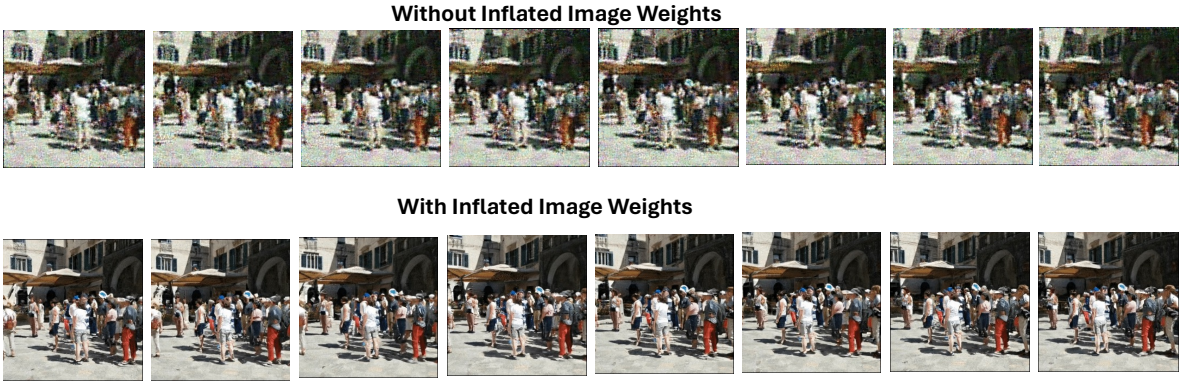
Figure 6: Text prompt: Little beautiful ducklings on green screen.

Figure 7: Text prompt: Brazil northeast beaches.

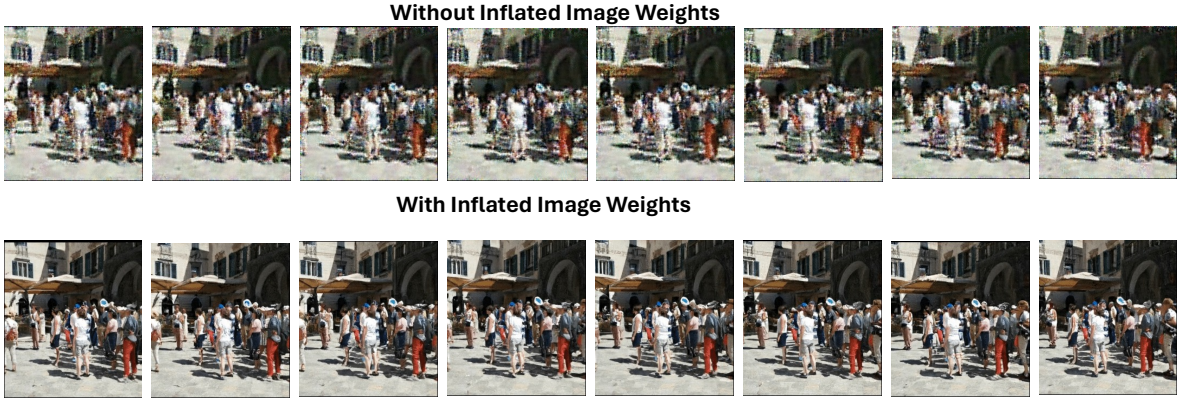
4.1. Quantitative Results

We evaluate different optimization approaches using various metrics. As shown in Table 1, the Full-ft approach achieves the best visual quality in terms of Peak signal to noise ratio (PSNR) and structural index similarity (SSIM) by tuning all 628.89 million parameters of UNet. The efficient temporal adapter tuning still yields reasonable visual quality while achieving an approximate $2\times$ wall-clock training acceleration and halving memory usage by adjusting only

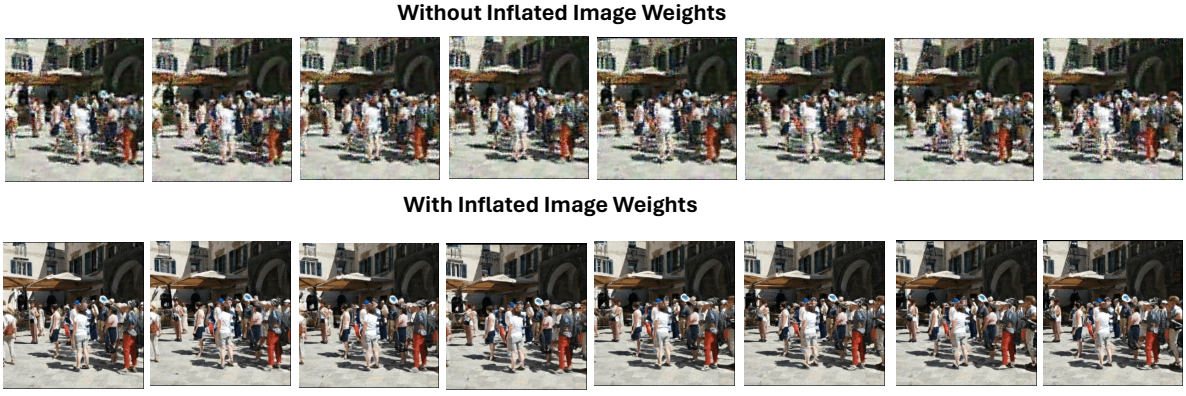
(a) Ground Truth



(b) Train with 40 % video data



(c) Train with 50 % video data



(d) Train with 60 % video data

Figure 8: Visualizations of methods with and without image model inflation. Text prompt: Tourists visiting the old town.

Table 1: Quantitative results for different tuning approaches.

Method	Visual Quality		Temporal Consistency	Efficiency		
	PSNR (% ↑)	SSIM (↑)	TCC (↑)	Tunable Params (M) ↓	Train Speed (steps/s) ↑	Memory (G) ↓
Zero-shot	18.1	0.42	0.70	-	-	-
Full-ft	28.7	0.77	0.86	628.89	1.05	15
Temporal	24.3	0.62	0.82	67.24	2.02	8

one-tenth of the typical parameter quantity. The zero-shot approach performs the worst.

We also validate that the efficient tuning approach can maintain temporal consistency, i.e. the motions among constructive frames remain smooth in the super-resolution results. We adopt the quantitative evaluation metric in [17]: temporal change consistency (TCC), which is defined as:

$$TCC(H, G) = \frac{\sum_{i=1}^{n-1} SSIM(|h^i - h^{i+1}|, |g^i - g^{i+1}|)}{n-1} \quad (2)$$

where $H = \{h^1, h^2, ..., h^n\}$ and $G = \{g^1, g^2, ..., g^n\}$ are high-resolution ground-truth and generated video frames, respectively. Table 1 shows a clear trade-off between training efficiency and temporal consistency, in which efficient temporal tuning still yields reasonable results. We also observe that zero-shot approach fails to maintain the consistent changes among adjacent frames due to the lack of a temporal module that operate exclusively on the time axis.

4.2. Qualitative Results

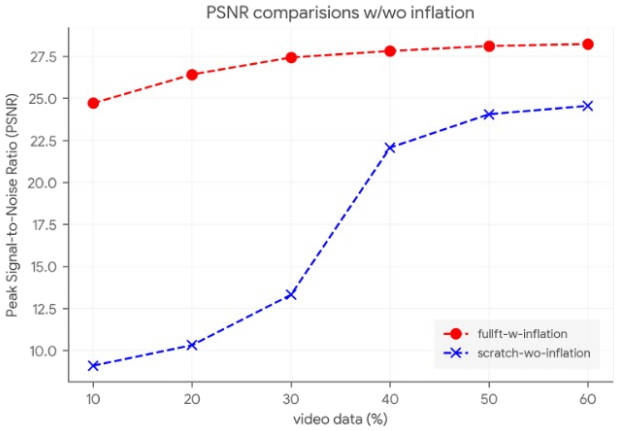
As shown in Figure 4, when compared with the ground truth high-resolution video (GT), both Full and Temporal produce good super-resolution results, marked by high visual quality and temporal smoothness. The ZS approach manages to generate frames with decent visual content without any fine-tuning on video data, but it falls short in maintaining temporal coherence — a limitation due to its pre-training solely on static images. This demonstrates the effectiveness of our temporal adapter in capturing the coherence across video frames. We provide addtional visualizations in Figure 5, 6 and 7.

4.3. Inflation is Data Efficient

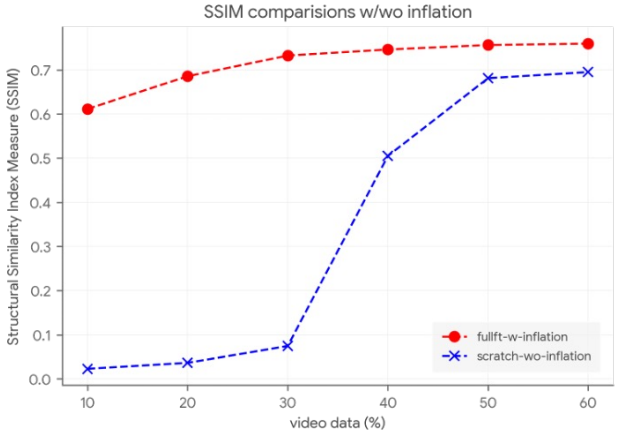
A straightforward baseline for image weight inflation is to randomly initialize the video UNet and fully fine-tune it using only video data. As observed in Figure 9, the image inflation-based approach can achieve high visual quality even when leveraging only 10% of 7M video data. This trend becomes more evident in Figure 8, demonstrating the data efficiency of the image weight inflation approach.

5. Conclusion

In this paper, we proposed a practical diffusion system for inflating text-to-image model weights to text-to-video spatial super-resolution model. This is the first work to study the weight inflation on the pixel level diffusion model. We have investigated different tuning methods for efficient temporal adaptation. We also demonstrated a good trade-off between the super-resolution quality with temporal consistency and tuning efficiency. As a future investigation, we aim to scale up our target resolution from 256 to 512 (e.g. from 4× to 8×SR) and generate videos with longer time frames, which would yield a more obvious trade-off between generation quality and computational resources.



(a) PSNR



(b) SSIM

Figure 9: Training data efficiency evaluated by PSNR and SSIM.

References

[1] 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 *CVPR*, 2023. 1, 2

[2] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. *CoRR*, abs/2303.12688, 2023. 1, 2

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[4] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022. 1, 2

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 3

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1

[7] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *CoRR*, abs/2303.13439, 2023. 1, 2

[8] 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. *J. Mach. Learn. Res.*, 2020. 3

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[10] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 2

[11] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer G. Dy and Andreas Krause, editors, *ICML*, 2018. 4

[12] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 1, 2

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[14] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1

[15] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *CoRR*, abs/2212.11565, 2022. 1, 2

[16] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 1

[17] Haokui Zhang, Ying Li, Yuanzhouhan Cao, Yu Liu, Chunhua Shen, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *ICCV*, 2019. 7

[18] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 1

[19] Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. Towards consistent video editing with text-to-image diffusion models. *CoRR*, 2023. 1, 2