Performance Comparison of Text-to-Video Generative Models

Gourav Kashiv[1], Sarbjeet Singh[2]

{gouravkashiv3@gmail.com[1], sarbjeet@pu.ac.in[2]}

University Institute of Engineering and Technology, Panjab University, Chandigarh, India[1,2]

Abstract. Text-to-video (T2V) generation has been significantly improved using diffusion-based generative models, but the generation of high-quality videos remains a difficult challenge. This paper examines how LoRA-FA can be used to fine-tune multiple T2V generative models, such as ModelScope, Zeroscope-320, and Zeroscope-576 with fewer parameters and under limited resources. Through quantitative and qualitative analysis we demonstrate that by applying Lora-FA, semantic alignment and perceptual sharpness can be improved, effectively increasing the quality of video frames generated. We found out that the Zeroscope v2(576w) model produced significant improvements in semantic alignment and perceptual sharpness while the improvements made by ModelScope were more balanced. The findings also indicate that LoRA-FA is useful for enhancing the quality of videos in low-resource configurations without having to retrain a whole system and lead the way to further studies of multimodal, domain-specific video generation tasks, which offers an efficient, scalable structure for effective T2V model optimization.

Keywords: Diffusion models, low-rank adaptation, large language models, feature alignment.

1 Introduction

The spread of generative artificial intelligence has democratized the creation of multimedia, making it possible for people to create realistic images and videos by simply typing a few words. Nevertheless, the task of video synthesis of text is extremely difficult, because it requires maintaining the consistency and quality of a series of frames over time. One of the most important weaknesses of the modern diffusion-based models is their tendency to generate blurry and repetitive sequences; the effect is caused by the fact that the textual prompt, although being a constraint, does not prevent the model to map specific noise patterns to visually similar results. Although the temperature parameter is used to control the local structure, it does not guarantee global temporal coherence and sharpness throughout the video. The development of photorealistic videos has been made possible through the use of transformers [12].

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362706

Large language models (LLMs) also have a lot of potential in generative systems, due to their large semantic knowledge and contextual reasoning capabilities [4]. Text-to-video (T2V) systems, a combination of LLMs and diffusion models, can further refine and align the complex textual and visual generations, and therefore improve the grounding of scenes and narrative consistency [5]. Although the large T2V models have sophisticated capabilities, the large T2V models are computationally intensive, thus limiting its use by researchers and practitioners [6].

In order to overcome the above challenges, parameter-efficient fine-tuning methods, including Low-Rank Adaptation (LoRA) [7] and its feature-aligned counterparts (LoRA-FA), have been proposed. LoRA-FA, with a computational cost of almost zero, can be used to improve the semantic fidelity and perceptual quality of pretrained models by modifying a small number of parameters [8]. Empirical studies found that LoRA-FA is always able to enhance the performance of text-to-video diffusion models and semantics are consistent across a variety of resolutions and architectural configurations.

LoRA-FA was used in the current research on the top text-to-video models, such as Model Scope [9], Zeroscope, and its various variations [23][24]. Our assessment system combines qualitative evaluation with quantitative measures, i.e., CLIP similarity [10], sharpness, SSIM [11], and PSNR [12], to measure semantic fidelity, consistency, and the quality of the video in general. We will use our experiments to show that LoRA-FA is a dual-efficacy method, both in terms of computational efficiency and quality improvement when used in video synthesis, thus making the method a viable solution to the high-quality, scalable problem of T2V synthesis.

To illustrate the difficulties that arise when considering all three challenges, namely semantic alignment, temporal coherence, and perceptual quality, this paper contributes

- the architectures, capabilities, and limitations of state-of-the-art T2V models [13][14].
- parameter-efficient fine-tuning through LoRA-FA to enhance video quality and maintain semantic consistency and lower computational requirements [8].
- model performance before and after fine-tuning, using both qualitative and quantitative metrics to objectively assess improvements and trade-offs in video generation quality.

2 Related Work

Text-to-video (T2V) generation is developed based on the improvement of diffusion-based frameworks and the corresponding models. Large language models (LLMs) are based on scalable attention models that are added to transformer networks [17]. These models take advantage of advanced reasoning and semantic representation functions, which are the foundation of multimodal tasks. The development of diffusion models, which are trained to produce both structured and unstructured data, such as stochastic noise, has also advanced the field, allowing the creation of visually plausible and temporally consistent videos [1], [18].

Based on these developments, a number of text-to-video generative models have been developed. An example of an open source solution is ModelScope T2V [9] that can generate short video clips with acceptable coherence and fidelity and is used by many. Zeroscope and its derivatives are more efficient in the training and also improve in temporal resolution and smoothness [23][24]. However, the field is still bound by trade-offs between the quality of output, the cost of computation and scalability.

T2V models are costly to fine-tune even though they have great potential. To alleviate this challenge, a parameter-efficient fine-tuning paradigm, which only trains a small number of parameters, has been put forward [6], [19]. Through an adaptive low-rank decomposition [7], light-weight learnable matrices are injected into attention layers, which significantly reduces the computational requirements. This approach can be extended to LoRA-FA [8], a method that allows to fine-tune without retraining the entire model. LoRA-FA achieves better semantic fidelity, perceptual clarity, and feature alignment, saving a lot of computational resources. This efficiency makes it feasible to adapt even large models to new tasks or domains.

This research aims at adapting existing text-to-video (T2V) models to new tasks and larger scopes. Evaluating these systems requires going beyond traditional metrics such as PSNR [20] and SSIM [11], which measure the structural fidelity and sharpness of the generated content, but fail to assess whether the generated video conveys the intended semantic content and narrative[12]. Based on this, a multimodal evaluation framework should be developed with multimodal measurements, such as CLIPScore [10], which measures the similarity between the textual and visual representations. The combination of high level semantic measures and the conventional low level measurements is a more comprehensive evaluation of the model performance.

Standardization in this area has seen efforts result in large benchmark suites including EvalCrafter [15], T2VBench [16] and FETV [21]. These suites provide integrated prompts, composition analysis software, and procedures of the gathering of human preference information. With the use of such benchmarks, researchers will be able to systematically determine the strong and weak aspects of individual models, not only in relation to semantic alignment, but also in relation to temporal continuity and visual realism. Additionally, common standards promote consistency in reporting, which enables making meaningful comparisons between models, baselines, and new approaches such as LoRA -FA and other strategies of fine-tuning.

Diffusion models, transformer-based architectures, and parameter-efficient fine-tuning (PEFT) have become the cornerstone methodologies in contemporary text-to-video (T2V) research. Such developments have produced high-quality gains in temporal coherence and semantic fidelity, especially when state-of-the-art models are coupled with lightweight fine-tuning techniques [22]. Empirical results suggest that parameter-efficient methods such as LoRA-FA can achieve high returns without the large-scale fine-tuning computational costs. However, there are still ongoing issues: the current systems still have the problem of motion instability, the proliferation of artifacts, and the inability to produce long and continuous video sequences. The further evolution of efficient and effective fine-tuning techniques implies a way out, the one that would balance the quality of

models with the realistic resource limitations, thus allowing the wider implementation of scalable T2V systems.

3 Fine-Tuning of Text-To-Video Models

This paper proposes a method to enhance T2V diffusion models by using the LoRA-FA architecture [8]. The first goal is to achieve strong adaptation of pre-trained models, including Text-to-Video MS-1.7B [9] and Zeroscope models [23][24] to a personalized dataset that is semantically stable and of high visual quality. The approach is based on the principles of transparency and reproducibility, which includes the choice of the model, data generation, fine-tuning, video generation, and evaluation.

To conduct comparative fine-tuning experiments, we chose three popular T2V models namely, Modelscope 1.7b, Zeroscope v1.1 320s, and Zeroscope v2 576w. All three of the models are based on diffusion architecture and have been scaled to our hardware, an NVIDIA T4 with 16GB of VRAM. The complementary architecture of these models offers a balanced range of evaluation of the flexibility of LoRA-FA in a wide range of model scales and capabilities.

A fine-tuning data set was compiled to help the models to learn meaningful relationships between text and visual data. This data consists of 1,970 short videos (16 frames or about two seconds) accompanied by natural-language prompts. This highly filtered data will enable the models to better project a broad set of textual descriptions to visual representations. All videos were downsampled to 256 256 resolution and the text prompts that accompanied them were preprocessed to fit the transformer encoder specifications of the models.

In order to guarantee strict evaluation, a second independent dataset of ten unseen prompts, which were semantically discontinuous was developed. Omission of these prompts in the fine-tuning data allows performance to be measured on the generalisation capacities of the models and not memorisation. This two-dataset design offers an objective and reliable foundation of assessing the performance of the models.

The model was fine-tuned on an NVIDIA T4 GPU. To operate within the 16GB VRAM constraint, we employed mixed-precision training (fp16) and gradient checkpointing, which were essential for managing memory usage The LoRA-FA architecture naturally minimizes the memory footprint by only updating a small portion of the model parameters, the low-rank matrices, as opposed to all of the billions of weights. This efficiency is fundamental to the method's low cost, allowing it to be run on limited hardware. The most critical hyperparameters for achieving an optimal balance between training efficiency and the expressiveness of the adaptation layers were a learning rate of 1×10^{-4} and a LoRA rank of 4.

The analysis used both quantitative and qualitative measures to provide an all-inclusive picture of the performance. Quantitative measurements were done using PSNR, SSIM, and CLIPScore to measure the quality of reconstruction, structural fidelity and semantic alignment respectively. Qualitative analysis was performed to assess aspects that are difficult to quantify, such as motion

realism, narrative coherence, and general visual plausibility. Workflow for Fine-Tuning and Evaluating Text-to-Video Models Using Lora-FA is shown Figure 1.

The pipeline discussed in steps is reproducible and demonstrates a viable way of parameter-efficient fine-tuning of T2V generation. Results show that LoRA-FA improves both visual and semantic fidelity of generated videos without retraining on a large scale, making large-scale video generative models more accessible to research and practice.

The fine-tuning process included the following steps. To begin with, a set of matched text-video clip pairs was carefully selected and an evaluation set with novel prompts was formed. The pre-trained models of ModelScope and Zeroscope were loaded and all the existing weights were frozen to maintain the pre-acquired knowledge. By using only lightweight LoRA and feature alignment layers, we kept the entire pre-trained backbone frozen. This eliminated the need for backpropagation through all model layers, leading to significantly faster training and reduced memory consumption.

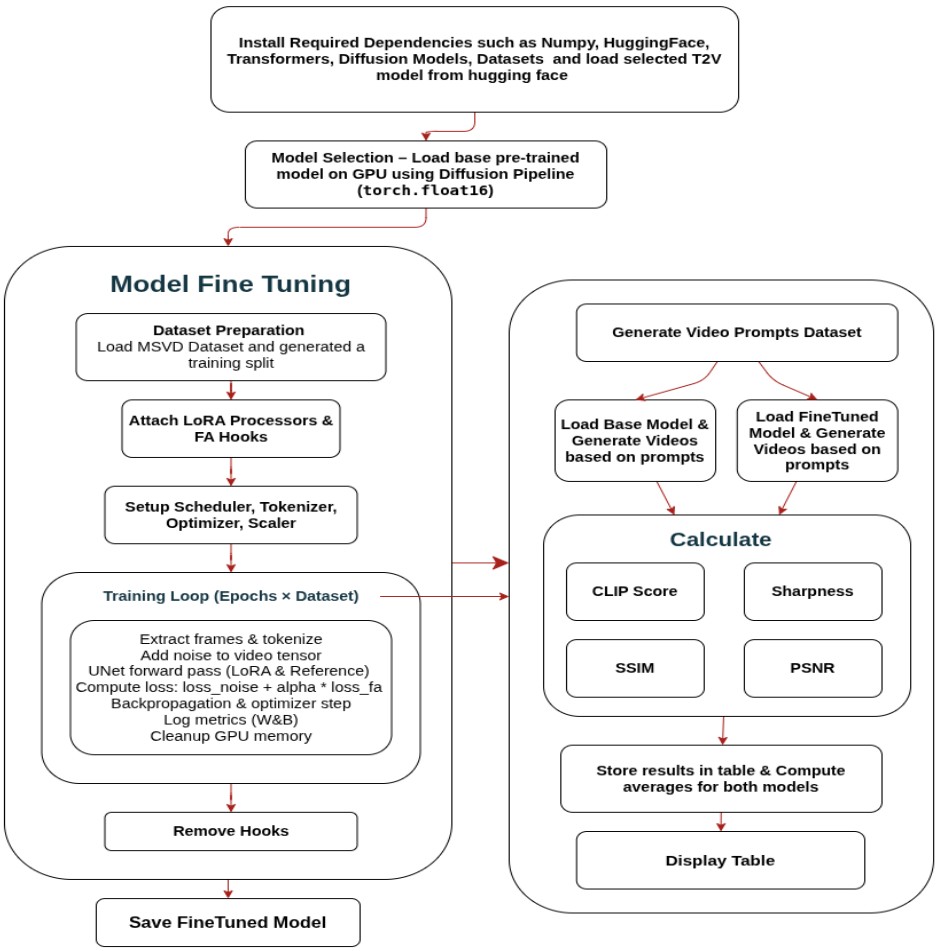


Fig.1. Workflow for Fine-Tuning and Evaluating Text-to-Video Models Using Lora-FA

One training epoch with a constant set of optimally selected hyperparameters was run on an NVIDIA T4. This one-epoch approach highlights the effectiveness of LoRA-FA, which can attain significant performance improvements with a small amount of computational cost, which is not possible with full-model fine-tuning using the same amount of resources. After the training, the results of the unchanged (base) and optimised (fine-tuned) models were obtained. The sets of resulting results were then evaluated systematically, and the effects of our methodology could be compared in detail.

In general, the paper describes a single methodological approach to using state-of-the-art text-to-video diffusion models through LoRA-FA fine-tuning. The framework is proven to reduce the hardware barrier significantly since all the experiments were run on a single mid-range GPU (NVIDIA T4). This illustrates that our methodology does not require high-cost and multi-GPU servers to work effectively in terms of text-to-video adaptation. The findings are made reliable and reproducible by rigidly outlining the training and evaluation processes. This systematic procedure will be further discussed in the next section of the article, where the experimental results and the performance study of the described pipeline on the whole will be offered.

4 Results and Evaluation

The current work used LoRA-FA fine-tuning on three text-to-video (T2V) diffusion models namely, ModelScope, Zeroscope v1.1 (320s), and Zeroscope v2 (576w). The assessment included quantitative measures (CLIP similarity scores, sharpness indices), as well as qualitative measures of semantic fidelity, temporal coherence, and visual artefacts. All the models were tested in their pretrained and fine-tuned versions with these evaluation measures.

The main goal of the current study was to identify the benefits brought by LoRA-FA fine-tuning for perceptual sharpness, semantic consistency, and adaptive ability for models of different architectural designs and image resolutions. Quantitative analyses showed a consistent improvement of all the models considered after the fine-tuning procedure. Notably, ModelScope, Zeroscope v1.1 (320 s) and Zeroscope v2 (576 w) showed improvements in their CLIP alignment scores of 1.8, 3.9 and 6.6 points respectively. These results indicate that LoRA - FA fine - tuning is especially effective for models with an initially lower performance, such as Zeroscope v2 (576w), while models with a higher baseline fidelity, such as ModelScope, showed smaller, but more consistent improvements.

The three models showed significant improvements in visual acuity, in terms of increased clarity and more detailed depiction of textures, in the generated video. Quantitatively, Zeroscope v1.1 (320 s) showed a significant increase in sharpness metric (7.8% i.e., from 924.50 to 997.03). ModelScope showed an even higher escalation with its score increasing by 18.3% (from 589.47 to 697.51). The biggest improvement was seen in Zeroscope v2 (576 w) with a 36.8% increase in the sharpness metric (from 601.92 to 823.69). These results highlight the potential of parameter-efficient fine tuning, especially in high-resolution architectures, to provide significant perceptual gains.

The differential adaptation among the models is further explained by analyzing the structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR) which are commonly used metrics for assessing the structural and signal fidelity. Zeroscope v1.1 (320 s) and Zeroscope v2 (576 w) are more out of line with their baseline outputs which is characteristic of intentional adaptation: the models apparently are trained to match more closely the semantics of the prompts than to reproduce original patterns. The performance summary by metric category is shown in Table 1. On the contrary, ModelScope showed a more balanced trade off, which is characterized by moderate structural

deviation. As such, the flexibility provided by LoRA-FA is dependent on the original capacity and resolution of the base model.

Table 1. Performance summary by metric category

Metric Category	ModelScope	Zeroscope v1.1 (320s)	Zeroscope v2 (576w)	Best Performer
Semantic Alignment (CLIP % ↑)	+1.8%	+3.9%	+6.6%	Zeroscope v2
Perceptual Quality (Sharpness % ↑)	+18.3%	+7.8%	+36.8%	Zeroscope v2
Structural Fidelity (SSIM)	0.2275	0.1548	0.1196	ModelScope
Signal Quality (PSNR)	13.01	11.30	8.57	ModelScope

These differences are further depicted by qualitative observations as opposed to quantitative measures. ModelScope had good temporal consistency but poor frame-to-frame coherence, which made it appropriate in the applications that required good sequential detail. Zeroscope v1.1 (320s) was better at object-interaction prompts, though sometimes at the cost of character detail (textures) in favor of finer detail. On the other hand, Zeroscope 2 (576w) had the most significant advances in scenes with complicated movements and detailed interactions. Such visual improvements especially in dynamic situations are indisputable. However, the adaptations in Zeroscope v2 (576w) can sometimes become aggressive resulting in slight artefacts like flickering or slight background discrepancies. Although minor problems with details, such as bleeding and smoothing of colours, were mostly reduced in the output of ModelScope, they were more noticeable in the Zeroscope models.

In general, the results indicate that LoRA-FA fine-tuning is a promising technique of improving the realism of motion, visual sharpness, and semantic consistency, but with relatively small increases in computational expenses. Nevertheless, the performance of the models is not uniform: Zeroscope v1.1 (320s) has the highest level of semantic alignment, ModelScope provides the most balanced performance, and Zeroscope v2 (576w) provides the most advanced results (perceptually) but with

structural instability in some cases. Taken together, these results suggest that LoRA-FA is a useful low-resource adaptation method to pretrained T2V diffusion models. Further improvements can include the use of larger, more diverse datasets, more advanced temporal optimization, or hybrid fine-tuning schemes, which could open the door to even greater performance and fidelity.

Although the extent and character of the improvement also depended on the architecture, quantitative and qualitative analysis prove that LoRA-FA always improves performance in all three models. The Consolidated comparison of the base and LoRA-FA fine-tuned text-to-video models is shown in Table 2. ModelScope had low losses in semantic consistency and perceptual acuity, and large temporal consistency and structural fidelity. Zeroscope v2 (576w) showed the most perceptual improvements in sharpness and motion intensive prompts even with relatively higher structural deviations, whilst Zeroscope v1.1 (320s) showed significant semantic alignment especially in object interaction prompts.

Table 2. Consolidated comparison of the base and LoRA-FA fine-tuned text-to-video models.

Model	CLIP Score (Base)	CLIP Score (FT)	CLIP % Improvement	Sharpness (Base)	Sharpness (FT)	Sharpness % Improvement
ModelScope	0.3216	0.3274	+1.8%	589.47	697.51	+18.3%
Zeroscope v1.1 (320s)	0.3088	0.3207	+3.9%	924.50	997.03	+7.8%
Zeroscope v2 (576w)	0.2888	0.3078	+6.6%	601.92	823.69	+36.8%

Lastly, the findings support the hypothesis that LoRA-FA is a practical and cost-efficient way of enhancing text-to-video models. Zeroscope v2 (576w) showed the largest perceptual improvement and highest semantic alignment. ModelScope found a compromise between the two dimensions. The result is that LoRA-FA fine-tuning is a cost-efficient and effective adaptation modality, which can significantly increase the quality of video generation across a range of architectures and resolutions. Minor artefacts, such as transient visual anomalies or temporal discrepancies, point to future improvements, such as augmented temporal regularization, larger training data, and the incorporation of complementary fine-tuning methods.

5 Conclusion, Discussion and Future Scope

This paper shows that Low-Rank Adaptation with Feature Alignment (LoRA-FA) is an effective method for parameter-efficient fine-tuning of text-to-video diffusion models, such as ModelScope, Zeroscope v1.1 (320s), and Zeroscope v2 (576w). The work presented in this paper significantly enhances semantic consistency and quality of perception and changes only a small part of model parameters.

The quantitative evaluation indicates that all three architectures showed measurable improvement. Specifically, Zeroscope v2 (576w) exhibited the strongest improvement in semantic alignment with a 6.6 per cent increase in CLIP score, and Zeroscope v1.1 (320s) and ModelScope have more moderate improvements of 3.9 per cent and 1.8 per cent, respectively. ModelScope has the most balanced performance in all metrics, which is retaining structural integrity and gaining an 18.3% in sharpness.

These results are backed by qualitative analysis finely tuned models are shown to produce videos with less motion flicker, increased coherence between frames and compliance with textual cues. Despite the presence of minor artifacts in complex sequences, the low-resource requirements of LoRA-FA make it an affordable and flexible solution to improve the quality of videos in a wide variety of model architectures.

Future work will be directed to three important aspects to further develop the solid foundation of LoRA-FA: First, diversifying fine-tuning datasets to improve the generalization ability of models; Second, addressing temporal abnormalities by recurrent architectures or motion-supervision schemes; Third, developing hybrid parameter-efficient algorithms to improve training stability. These directions are aimed at making more powerful and accessible deep video synthesis technologies.

References

[1] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)

[2] Ho, J., et al.: Imagen Video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)

[3] Blattmann, A., et al.: Align your latents: High-resolution video synthesis with latent diffusion models. arXiv preprint arXiv:2304.08818 (2023)

[4] Lian, L., Shi, B., Yala, A., Darrell, T., Li, B.: LLM-grounded video diffusion models. arXiv preprint arXiv:2309.17444 (2023)

[5] Yin, A., et al.: The best of both worlds: Integrating language models and diffusion models for video generation. arXiv preprint arXiv:2503.04606 (2025)

[6] Xu, L., Xie, H., Qin, S.Z.J., Tao, X., Wang, F.L.: Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. arXiv preprint arXiv:2312.12148 (2023)

[7] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR) (2022)
[8] Zhang, L., Zhang, L., Shi, S., Chu, X., Li, B.: LoRA-FA: Memory-efficient low-rank adaptation for large language models fine-tuning. arXiv preprint arXiv:2308.03303 (2023)
[9] Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: ModelScope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
[10] Hessel, J., Holtzman, A., Forbes, M., le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
[11] Nilsson, J., Akenine-Möller, T.: Understanding SSIM. arXiv preprint arXiv:2006.13846 (2020)
[12] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing. Vol. 13, no. 4, pp. 600–612 (2004)
[13] Ma, X., et al.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
[14] Bao, F., et al.: Vidu: A highly consistent, dynamic and skilled text-to-video generator with diffusion models. arXiv preprint arXiv:2405.04233 (2024)
[15] Liu, Y., et al.: EvalCrafter: Benchmarking and evaluating large video generation models. arXiv preprint arXiv:2310.11440 (2023)
[16] Li, X., et al.: T2VBench: Benchmarking temporal dynamics for text-to-video generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (2024)
[17] Vaswani, A., et al.: Attention is all you need. Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
[18] Wang, Y., et al.: Survey of video diffusion models: Foundations, implementations, and applications. arXiv preprint arXiv:2504.16081 (2025)
[19] Wang, L., et al.: Parameter-efficient fine-tuning in large language models: A survey of methodologies. Artificial Intelligence Review. (2025)
[20] Fardo, F.A., Conforto, V.H., De Oliveira, F.C., Rodrigues, P.S.: A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms. arXiv preprint arXiv:1605.07116 (2016)
[21] FETV: Fine-grained evaluation of text-to-video generation. Advances in Neural Information Processing Systems (NeurIPS). (2023)
[22] Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7623–7633 (2023)
[24] Cerspense: zeroscope_v1-1_320s. Hugging Face Repository, https://huggingface.co/cerspense/zeroscope_v1-1_320s (2023)