Multimodal Generation with Consistency Transferring

Junxiang Qiu, Jinda Lu*, Shuo Wang
University of Science and Technology of China
qiujx@mail.ustc.edu.cn, lujd@mail.ustc.edu.cn, shuowang.edu@gmail.com

Abstract

Multimodal content generation has become an area of considerable interest. However, existing methods are hindered by limitations related to model constraints and training strategies: (1) Most current approaches rely on training models from scratch, resulting in inefficient training processes when extending these models; (2) There is a lack of constraints on adjacent steps within the models, leading to slow sampling and poor generation stability across various sampling methods. To address the issues, we introduce Multimodal Generation with Consistency Transferring (MGCT). The method introduces two key improvements: (1) A Model Consistency Transferring (MCT) strategy to acquire low-cost prior knowledge, increasing training efficiency and avoiding error accumulation; (2) A Layer Consistency Transferring (LCT) between adjacent steps, enhancing denoising capabilities at each step and improving model stability across various generation methods. These strategies ensure the consistency of jointly generated multimodal content and improving training efficiency. Experiments show that the algorithm enhances the model's ability to capture actions and depict backgrounds more effectively. In both the AIST++ and Landscape datasets, it improves video generation speed by approximately 40% and quality by about 39.3%, while also achieving a slight 3% improvement in audio quality over the baseline.

1 Introduction

In recent years, content generation in the fields of image ((Rombach et al., 2022; Saharia et al., 2022a; Ramesh et al., 2022; Chang et al., 2023)), video ((Blattmann et al., 2023; Singer et al., 2022; Guo et al., 2019; Wang et al., 2018; Hao et al., 2022)), and audio ((Huang et al., 2023)) has garnered significant attention from both the global academic community and industry. However, most image and video generation methods are limited to unimodal content production. Existing multimodal approaches ((Ruan et al., 2023; Tang et al., 2024; Wang et al., 2019)) for generating audio and video face several challenges, including high training complexity, significant costs, heavy reliance on dataset quality, and inadequate emphasis on model constraints and training strategies.

Using diffusion models ((Ho et al., 2020; Sohl-Dickstein et al., 2015)), significant progress has been made in image generation, video generation, and multimodal generation. Some recent methods ((Ruan et al., 2023)) have successfully achieved joint audio-video generation. However, unlike image and short video generation, joint audio-video generation presents higher complexity and demands more sampling steps, resulting in longer sampling times. However, continuing to use the common approach of training multimodal generation models from scratch to address the aforementioned issues may lead to error accumulation and inefficiency. To mitigate this issue, we employ Model Consistency Transferring (MCT) strategy, which leverages the initial training methods of MM-Diffusion (Ruan et al., 2023). This allows for the rapid acquisition of preliminary knowledge about the generation scene at a low cost, significantly improving training efficiency.

Furthermore, it has been observed that maintaining the quality and stability of generated samples across different sampling methods, such as DDPM ((Ho et al., 2020)) and DPM-Solver ((Lu et al., 2022)), is challenging with existing joint audio-video generation techniques. These current methods often fail to consider the relationships between adjacent steps in diffusion models, suggesting that there is still considerable room for improvement in both generation speed and stability.

The introduction of the Consistency Model ((Song et al., 2023)) offers a promising strategy to mitigate the high sampling costs associated with

[*] Jinda Lu is the corresponding author

Findings of the Association for Computational Linguistics: NAACL 2025, pages 504–513
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

diffusion models and enhance the sampling effectiveness at each step. This model enhances the denoising effectiveness at each step, reducing the number of sampling steps needed for image synthesis, and thereby accelerating the generation process and improving generation stability. While this approach has been successfully applied in image generation ((Luo et al., 2023a)) and video generation ((Wang et al., 2023)), its potential in joint audio-video generation has not yet been explored. To address this gap, we propose a cross-modal Layer Consistency Transferring (LCT) method that applies the Consistency Model to enforce consistency across adjacent steps in audio-video sampling.

Therefore, in this paper, we introduce Multimodal Generation with Consistency Transferring (MGCT) method, which improves the training efficiency. Specifically: we employ MCT strategy, continuing to use the MM-Diffusion training method in the early stages of training, to quickly learn preliminary knowledge corresponding to the generation scenarios under low-cost conditions. At the same time, we have proposed a cross-modal LCT method that enforces consistency constraints on the sampling capability of the consistency model for adjacent steps in audio-visual sequences. In summary, the contributions of MGCT are threefolds:

- We introduced Model Consistency Transferring, which reduces the training time cost of the model and effectively avoids the accumulation of errors.
- We designed Layer Consistency Transferring to enforce consistency in adjacent frame generation within the model, thereby enhancing its denoising capability.
- In the experiments, the model is able to generate high-quality audio and video content in a shorter amount of time. Compared to the slow sampling method, the fast sampling method can generate samples of equivalent quality.

2 Related Work

This paper primarily addresses Multimodal Diffusion Models and Consistency Models. This chapter offers a comprehensive overview of their background, highlights the characteristics and limitations of previous research, and details the enhancement methods we have selected.

Multimodal Diffusion Models. Probability diffusion models (DPMs) ((Ho et al., 2020; Sohl-Dickstein et al., 2015)) are a type of generative model that creates data samples from random noise. Compared to models like GANs ((Goodfellow et al., 2014)), DPMs deliver superior performance in terms of generation quality and diversity ((Dhariwal and Nichol, 2021)). They excel in various image generation tasks, including image inpainting ((Lugmayr et al., 2022)), image classification (Wang et al., 2022; Lu et al., 2023; Zhu et al., 2024; Wang et al., 2024a,b), super-resolution ((Saharia et al., 2022b)), textual generation (Ben et al., 2024), and image restoration ((Kawar et al., 2022)). However, DPMs typically suffer from slower sampling speeds due to the repeated denoising required during sampling. To make DPMs more practical, DDIM ((Song et al., 2020)) modifies the traditional diffusion model's reverse process, reducing the number of inference steps needed to generate images and thereby accelerating the sampling process. DPM-Solver ((Lu et al., 2022)) takes this further by solving the ordinary differential equations of the DPM reverse process and providing higher-order approximate solutions, significantly reducing the steps required for sampling.

As the theory of DPMs has evolved and been refined, diffusion models have increasingly been applied in the field of multimodal generation, such as text-to-vision ((Mou et al., 2024)), text-to-audio ((Liu et al., 2023)), vision-to-audio ((Luo et al., 2024)), and vision editing ((Ceylan et al., 2023)). Despite this progress, these approaches typically generate only one modality at a time. Some research has started to explore multimodal joint generation ((Ruan et al., 2023; Tang et al., 2024; Zhu et al., 2023)). MM-Diffusion ((Ruan et al., 2023)) represents a pioneering approach for simultaneous audio and video generation; however, it lacks constraints on the diffusion steps, which limits its scalability and effectiveness.

Consistency Models. To tackle the issue of slow generation speeds caused by the extensive inference steps required in diffusion models, consistency models ((Song et al., 2023)) were introduced. Built on the foundation of the probability flow ordinary differential equation (PF-ODE), consistency models are designed to map any point at any time step directly back to the start of the trajectory. Consistency models enable efficient one-step image generation without sacrificing the advantages of multistep iterative sampling, thereby achieving higher-quality results through multistep inference. Based on this, LCM ((Luo et al., 2023a)) explored

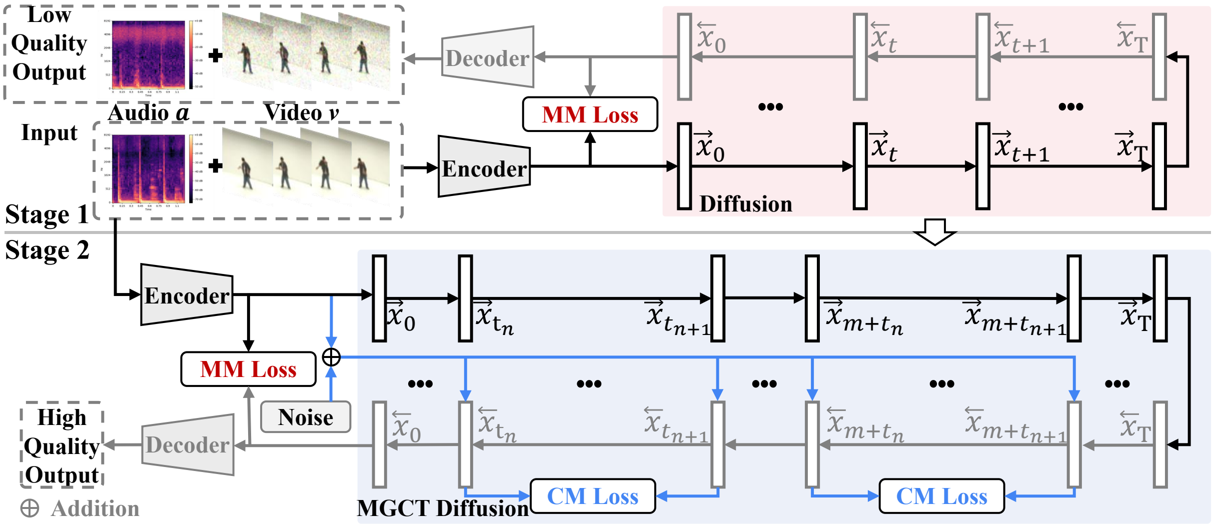

Figure 1: Overview of MGCT. In the Stage 1, we use MM-Diffusion for training to obtain initial weights with preliminary knowledge, and these weights serve as the starting model for our approach. Next, we transfer the prior knowledge from this model to the MGCT model, using it as the starting point of training in the Stage 2. During subsequent training, we retain the loss function between the ground truth and the reverse output, and incorporate a consistency loss to train the model on adjacent steps at fixed intervals, enhancing its denoising capability.

consistency models in latent space to reduce memory consumption and enhance inference efficiency. Subsequently, several methods ((Luo et al., 2023b; Xiao et al., 2023)) investigated efficient generation techniques and achieved significant results. Videolcm ((Wang et al., 2023)) extended consistency models to the video generation domain, reducing the sampling steps to single digits. Inspired by this, we further extend consistency models to the domain of audio-visual generation.

Based on the analysis of related work, the methods most relevant to MGCT are MM-Diffusion and Consistency Models. Recent multimodal generation techniques often focus on enhancing model functionality while neglecting the constraints between adjacent steps in diffusion models. This oversight can lead to insufficient denoising capabilities and a loss of generation quality, particularly when using fast sampling methods. Furthermore, the exploration of consistency models is still in progress. In MGCT, we extend the application of consistency models to the audio and video generation domains, applying constraints to adjacent frames to enhance denoising performance. Additionally, because consistency constraints can introduce greater convergence challenges and are prone to error accumulation, MGCT integrates MCT to address these issues.

3 Approach

Our approach incorporates consistency constraints between adjacent steps in multimodal generative models to address issues such as poor denoising performance at each step and unstable denoising results. By leveraging pretrained weights as a starting point, we streamline the learning process, accelerate training, and avoid error accumulation.

3.1 Preliminaries

The training data includes pairs of videos and audio, which we define separately as $\boldsymbol{v}$ and $\boldsymbol{a}$. The corresponding video $\boldsymbol{v}$ and audio $\boldsymbol{a}$ are collectively referred to as data $\boldsymbol{x}$. The overview of our approach is illustrated in Figure 1. Initially, we use the training methodology from MM-Diffusion to train an initial set of weights, which we designate as the start point of Stage 2. These weights are able to learn from the prior knowledge of the datasets, thus significantly dorping the MGCT training difficulty and reducing the time required for training. During the subsequent training phase (Stage 2), we retain the loss function that compares the ground truth with the output of the Diffusion Model. To further enhance this, we introduce an additional multimodal consistency loss. This consistency loss involves calculating the difference between the denoising results of consecutive frames, $\overleftarrow{\boldsymbol{x}}_{t_{n+1}}$ and $\hat{\boldsymbol{x}}_{t_n}$, for both the audio and video pairs. Finally, we combine the consistency loss and the multimodal loss by weighting them appropriately. By incorporating this consistency loss, we ensure that the model maintains consistent denoising capability at fixed intervals, thereby improving its overall denoising performance.

3.2 Layer Consistency Transferring

Based on diffusion models refer to a class of generative methods that first transform a given data distribution $\boldsymbol{x}$ into Gaussian noise (forward process) and then learn to recover the data distribution by reversing the aforementioned forward process (reverse process). In the following text, at step t, we refer to the data as $\boldsymbol{v}_t$, $\boldsymbol{a}_t$, and $\boldsymbol{x}_t$ respectively. And noise is progressively added to the data through a forward process, which is represented as follow:

$$\begin{aligned} &\boldsymbol{q}(\overrightarrow{\boldsymbol{x}}_{\boldsymbol{t}} \mid \overrightarrow{\boldsymbol{x}}_{\boldsymbol{t}-1}) = \\ &\mathcal{N}(\overrightarrow{\boldsymbol{x}}_{\boldsymbol{t}}; \sqrt{1-\boldsymbol{\beta_t}}\overrightarrow{\boldsymbol{x}}_{\boldsymbol{t}-1}, \boldsymbol{\beta_t} I), \end{aligned} \tag{1}$$

where $\overrightarrow{x}$ represents that x is in the forward process, $t \in [1, T]$, and $\boldsymbol{\beta_t}$ is a parameter associated with the time step. The entire process is modelled using Markov chains $\mathcal{N}$ starting at $p\left(\overleftarrow{\boldsymbol{x}}_T\right) = \mathcal{N}\left(\overleftarrow{\boldsymbol{x}}_T; 0, I\right)$, where $\overleftarrow{x}$ represents that x is in the reverse process (Similarly, $\overleftarrow{a}$ and $\overleftarrow{v}$ represent the audio and video data in the reverse process, respectively). Subsequently, the reverse process begins from the final noise state and progressively restores the data, represented as follow:

$$\begin{aligned} &p_{\boldsymbol{\theta}}(\overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}-1} \mid \overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}}) = \\ &\mathcal{N}(\overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}-1}; \boldsymbol{\mu_\theta}(\overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}}, \boldsymbol{t}), \Sigma_{\boldsymbol{\theta}}(\overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}}, \boldsymbol{t})). \end{aligned} \tag{2}$$

where μ_θ denotes the Gaussian mean value predicted by θ.

For multimodal generation, the forward process uses a noise-adding method similar to DDPM and processes audio and video separately. Therefore, this part is intentionally omitted. Unlike the forward process of independently modeling audio and video, we consider the correlation between the two modalities and propose a unified model $\boldsymbol{\theta_{av}}$, which takes both modalities as inputs and enhances each other's audio and video generation quality. The reverse process is as follows:

$$\begin{aligned} &\boldsymbol{p_{\theta_{av}}}(\overleftarrow{\boldsymbol{a}}_{\boldsymbol{t}-1} | (\overleftarrow{\boldsymbol{a}}_{\boldsymbol{t}}, \overleftarrow{\boldsymbol{v}}_{\boldsymbol{t}})) = \\ &\mathcal{N}(\overleftarrow{\boldsymbol{a}}_{\boldsymbol{t}-1}; \boldsymbol{\mu_{\theta_{av}}}(\overleftarrow{\boldsymbol{a}}_{\boldsymbol{t}}, \overleftarrow{\boldsymbol{v}}_{\boldsymbol{t}}, \boldsymbol{t})), \end{aligned} \tag{3}$$

where $\overleftarrow{\boldsymbol{a}}_{\boldsymbol{t}-1}$ is generated from a Gaussian distribution determined jointly by $\overleftarrow{\boldsymbol{a}}_{\boldsymbol{t}}$ and $\overleftarrow{\boldsymbol{v}}_{\boldsymbol{t}}$. The authors use $\boldsymbol{\epsilon}$-prediction, which is defined as follows:

$$\begin{aligned} &\boldsymbol{\mathcal{L}}_{MM} = \\ &\boldsymbol{E}_{\boldsymbol{\epsilon}\sim\boldsymbol{\mathcal{N}}_{\boldsymbol{a}(0,I)}}\left[\boldsymbol{\lambda}(t)\left\|\tilde{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}\left(\overleftarrow{\boldsymbol{a}}_{\boldsymbol{t}}, \overleftarrow{\boldsymbol{v}}_{\boldsymbol{t}}, \boldsymbol{t}\right) - \boldsymbol{\epsilon}\right\|_2^2\right], \end{aligned} \tag{4}$$

where $t \in [0, T]$, and $\boldsymbol{\lambda}$ is an optional weighting function. Video and audio have similar representations in this process.

In a diffusion model, generation is performed through multi-step sampling, whereas a consistency model assumes the existence of a function $\boldsymbol{f}$ that outputs the same value at each node in the aforementioned process, defined as follows:

$$\boldsymbol{f}\left(\overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}}, \boldsymbol{t}\right) = \boldsymbol{f}\left(\overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}'}, \boldsymbol{t}'\right), \tag{5}$$

where $\boldsymbol{t}, \boldsymbol{t'} \in [\alpha, T]$ and for the initial point of the trajectory $\overleftarrow{\boldsymbol{x}}_0 = \boldsymbol{\alpha}$, the following equation holds:

$$\boldsymbol{f}\left(\overleftarrow{\boldsymbol{x}}_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}\right) = \overleftarrow{\boldsymbol{x}}_{\boldsymbol{\alpha}}. \tag{6}$$

Then, for any point on the trajectory, substituting the prior distribution completes a one-step sampling. A neural network is trained to fit $\boldsymbol{f}$, but two conditions must be met: first, the output value must be consistent for points on the trajectory, and second, at the initial time point, $\boldsymbol{f}$ must be an identity function with respect to $\boldsymbol{x}$. Therefore, the following formula is designed:

$$\boldsymbol{f_\theta}\left(\overleftarrow{\boldsymbol{x}}, \boldsymbol{t}\right) = \boldsymbol{c}_{\text{skip}}\left(\boldsymbol{t}\right)\overleftarrow{\boldsymbol{x}} + \boldsymbol{c}_{\text{out}}\left(\boldsymbol{t}\right)\boldsymbol{F_\theta}\left(\overleftarrow{\boldsymbol{x}}, \boldsymbol{t}\right), \tag{7}$$

where $\boldsymbol{c}_{skip}$ and $\boldsymbol{c}_{out}$ are differentiable functions that satisfy $\boldsymbol{c}_{skip}\left(\boldsymbol{\alpha}\right) = 1$ and $\boldsymbol{c}_{out}\left(\boldsymbol{\alpha}\right) = 0$. $\boldsymbol{F_\theta}$ is a deep neural network with an output dimension equal to that of $\overleftarrow{\boldsymbol{x}}$. This ensures that:

$$\boldsymbol{f_\theta}\left(\overleftarrow{\boldsymbol{x}}, \boldsymbol{t}\right) = \begin{cases} \overleftarrow{\boldsymbol{x}}, & \boldsymbol{t} = \boldsymbol{\alpha}, \\ \boldsymbol{F_\theta}\left(\overleftarrow{\boldsymbol{x}}, \boldsymbol{t}\right), & \boldsymbol{t} \in (\boldsymbol{\alpha}, T]. \end{cases} \tag{8}$$

Previous works did not impose constraints between adjacent steps, leading to poor denoising capability at each step. This is the main reason for the longer time required for generation and the degradation in quality under fast generation methods. Therefore, we plan to use a consistency loss to constrain adjacent steps, ensuring that they have the same denoising capability. In this way, the same generation quality as MM-Diffusion can be achieved with fewer sampling steps. However, consistency models are known to be difficult to train, especially in the final stages, which require a significant amount of time. Thus, in this work, we only use it to improve the denoising capability of each step. The loss function of the consistency model is defined as follows:

$$\boldsymbol{\mathcal{L}}_{CD}^{N}\left(\theta,\theta^{-};\phi\right) := \boldsymbol{E}\left[\lambda\left(\boldsymbol{t}_n\right)\cdot \boldsymbol{d}\left(\boldsymbol{f}_{\boldsymbol{\theta}}\left(\overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}_{n+1}},\boldsymbol{t}_{n+1}\right),\boldsymbol{f}_{\boldsymbol{\theta}^{-}}\left(\hat{\boldsymbol{x}}_{\boldsymbol{t}_n}^{\phi},\boldsymbol{t}_n\right)\right)\right]. \tag{9}$$

$N-1$ represents the number of sub-intervals. And the expectation is taken with respect to $\boldsymbol{x} \sim p_{\text{data}}$, $n \sim u[1, N-1]$, and $\boldsymbol{x}_{\boldsymbol{t}_{n+1}} \sim \mathcal{N}\left(\boldsymbol{x};\boldsymbol{t}_{n+1}^{2}I\right)$. Here, $u[1, N-1]$ denotes the uniform distribution over $\{1, 2, \ldots, N-1\}$, $\boldsymbol{\theta}^{-}$ represents the running average of past $\boldsymbol{\theta}$ values during the optimization process, and $\boldsymbol{d}(\cdot,\cdot)$ is a metric function satisfying $\forall j, k : \boldsymbol{d}(j, k) \geq 0$ and $\boldsymbol{d}(x, y) = 0$ if and only if $j = k$. The definition of $\hat{\boldsymbol{x}}_{\boldsymbol{t}_n}^{\phi}$ is as follows:

$$\hat{\boldsymbol{x}}_{\boldsymbol{t}_n}^{\phi} := \overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}_{n+1}} + \left(\boldsymbol{t}_n - \boldsymbol{t}_{n+1}\right)\boldsymbol{\Phi}\left(\overleftarrow{\boldsymbol{x}}_{\boldsymbol{t}_{n+1}},\boldsymbol{t}_{n+1};\phi\right), \tag{10}$$

where $\boldsymbol{\Phi}(\cdots;\phi)$ denotes the update function applied to the empirical PF-ODE as a one-step ODE solver. Rewrite Equation 9 into multimodal form, the CM loss can be shown in Equation 11:

$$\boldsymbol{\mathcal{L}}_{CM}^{N}\left(\theta,\theta^{-};\phi\right) := \boldsymbol{E}\left[\lambda\left(\boldsymbol{t}_n\right)\boldsymbol{d}\left(\boldsymbol{f}_{\boldsymbol{\theta}}\left(\overleftarrow{\boldsymbol{a}}_{\boldsymbol{t}_{n+1}},\overleftarrow{\boldsymbol{v}}_{\boldsymbol{t}_{n+1}},\boldsymbol{t}_{n+1}\right),\boldsymbol{f}_{\boldsymbol{\theta}^{-}}\left(\hat{\boldsymbol{a}}_{\boldsymbol{t}_n}^{\phi},\hat{\boldsymbol{v}}_{\boldsymbol{t}_n}^{\phi},\boldsymbol{t}_n\right)\right)\right], \tag{11}$$

where $\hat{\boldsymbol{a}}_{\boldsymbol{t}_n}^{\phi}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{t}_n}^{\phi}$ are the audio and video representations of $\hat{\boldsymbol{x}}_{\boldsymbol{t}_n}^{\phi}$, respectively. As shown in Figure 1, MGCT builds upon MM-Diffusion by introducing the concept of a consistency model and incorporating a consistency loss $\boldsymbol{\mathcal{L}}_{CM}^{N}$ to focus on and unify the denoising capabilities of adjacent steps, thereby enhancing the denoising ability at each step. In terms of computation, we retain the MM loss (multimodal loss). Meanwhile, both $\boldsymbol{\mathcal{L}}_{CM}^{N}$ and $\boldsymbol{\mathcal{L}}_{MM}$ are calculated, and their weighted sum is computed. During training, we minimize the objective by applying stochastic gradient descent to the model parameters θ, and update θ^{-} using exponential moving averages (EMA). The MGCT training method is detailed in **Algorithm** 1.

Algorithm 1 MGCT Training

```
Input: dataset $\boldsymbol{x}$, initial model parameters $\boldsymbol{\theta}$, learning rate $\eta$, loss function ratio $w$, ema decay rate schedule $\mu(\cdot)$ and $\lambda(\cdot)$
$\boldsymbol{\theta}^{-} \leftarrow \boldsymbol{\theta}$
sample $\boldsymbol{a} \sim \boldsymbol{x}['audio']$, $\boldsymbol{v} \sim \boldsymbol{x}['video']$
while $\boldsymbol{\mathcal{L}} >= \boldsymbol{\delta}$ do
    $\boldsymbol{\mathcal{L}} \leftarrow \boldsymbol{\mathcal{L}}_{MM}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}\boldsymbol{\mathcal{L}}$
end while
while not converged and $\boldsymbol{\mathcal{L}} < \boldsymbol{\delta}$ do
    $\boldsymbol{\mathcal{L}} \leftarrow (1-\omega)\,\boldsymbol{\mathcal{L}}_{MM} + \omega\boldsymbol{\mathcal{L}}_{CM}^{N}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}\boldsymbol{\mathcal{L}}$
    $\boldsymbol{\theta}^{-} \leftarrow stopgrad\left(\mu(k)\,\boldsymbol{\theta}^{-} + (1-\mu(k))\,\boldsymbol{\theta}\right)$
end while
```

3.3 Model Consistency Transferring

Although consistency models can enhance the model's denoising capability, ensuring information consistency across different modalities requires the introduction of complex loss functions and training mechanisms, which increases the model's computational complexity and training time. Therefore, as shown in Figure 1, we initially use only multimodal loss during the early stages of model training, allowing the model to learn basic generative information. Then, after the training reaches a certain point, we transfer the pretrained model to a MGCT model of the same size. On this basis, we introduce the consistency model, allowing the two loss functions to work together. Specifically, we set a threshold δ to evaluate the degree of convergence. In these two stages, we have the Loss Function as shown in Equation 12 and Algorithm 1:

$$\boldsymbol{\mathcal{L}} = \begin{cases} \boldsymbol{\mathcal{L}}_{MM}, & \boldsymbol{\mathcal{L}} > \delta, \\ (1-\omega)\boldsymbol{\mathcal{L}}_{MM} + \omega\boldsymbol{\mathcal{L}}_{CM}^{N}, & \boldsymbol{\mathcal{L}} < \delta. \end{cases} \tag{12}$$

4 Experiments

In this section, we evaluate the proposed MGCT diffusion model and compare its joint audio and video generation performance with MM-Diffusion. Additionally, we conducte ablation experiments to verify the effectiveness of the method proposed in this paper. Our experiments are intended to address the following research questions (RQ):

RQ1: How does MGCT perform compared to the baseline methods?

RQ2: How does the loss function ratio affect the quality of generated samples?

RQ3: What are the effects of MCT?
RQ4: Quality of audio and video samples generated by MGCT?

4.1 Experiment Settings

In this experiment, we selected the AIST++ ((Lee et al., 2022)) and landscape ((Li et al., 2021)) datasets for testing, where AIST++ has high requirements for capturing the detailed contours and audio rhythm of characters, while landscape poses a challenge to the model's ability to generate large-scale landscapes. For the diffusion model, we set the diffusion step length T to 1000, and to accelerate the sampling process, we default to using DPM-Solver [27] unless additional annotations are provided. For video quality assessment, we employed the FVD and KVD metrics, and for audio quality evaluation, we used the FAD metric. When comparing different models, we randomly generated 2,048 samples for objective evaluation, and all calculations were performed at a resolution of 64×64. During the experiment, we utilized two A40 GPUs to support the computational demands.

4.2 Comparison with Other Methods

Comparison with baseline (RQ1). To verify the denoising capability of MGCT and its stability under different sampling methods, experiments in Table 1 and Table 2 use the slow generation method DDPM and the fast generation method DPM-Solver for sampling, and tests are conducted on the AIST++ and Landscape datasets. MM-Diffusion is used as the baseline for comparison.

We are able to find from Table 1 that: (1) In the comparison of DPM-Solver generation methods between MGCT and MM-Diffusion, MGCT significantly outperforms MM-Diffusion in both the AIST++ and Landscape datasets. For instance, on the AIST++ dataset, MGCT achieves the lowest FVD score of 209.729 with 20 inference steps, compared to MM-Diffusion's FVD of 345.724 under the same conditions. Similarly, MGCT performs better on the Landscape dataset, with an FVD of 313.545 compared to MM-Diffusion's 369.379 at 20 steps. This demonstrates MGCT's strong advantage in the domain of fast generation, particularly in scenarios where a reduction in inference steps is required while still maintaining high-quality output; (2) Regarding the impact of reducing inference steps on generation quality, MGCT demonstrates that even with only 12 steps of DPM-Solver, it can achieve or exceed the generation quality of MM-Diffusion using 20 steps. Specifically, MGCT achieves an FVD of 279.384 at 12 steps on the AIST++ dataset, which outperforms MM-Diffusion's FVD of 345.724 at 20 steps. This showcases MGCT's ability to maintain high performance with fewer inference steps.

From Table 2, it can be seen that using DDPM sampling, the generation quality of the Landscape dataset is superior to that of MM-Diffusion. For example, MGCT has an FVD of 330.745 and an FAD of 1.489, which are better than MM-Diffusion's FVD of 353.283 and FAD of 1.521. For the AIST++ dataset, MGCT also generates samples of comparable quality to MM-Diffusion.

Comparing the results from Table 1 and Table 2, we find that: (1) When comparing the DDPM generation methods of MGCT and MM-Diffusion, MGCT's DPM-Solver method in the AIST++ dataset produces samples of comparable quality to those generated by MM-Diffusion's DDPM method. This indicates that MGCT can maintain high generation quality even with fast generation methods. In the Landscape dataset, MGCT's DPM-Solver method even surpasses MM-Diffusion's DDPM method, achieving an FVD of 313.545, a KVD of 7.457, and an FAD of 1.216 with 20 inference steps, compared to MM-Diffusion's FVD of 353.283, KVD of 9.030, and FAD of 1.521. This highlights the superior balance MGCT achieves between generation quality and efficiency; (2) In terms of the generation quality difference between MM-Diffusion and MGCT, particularly in the AIST++ dataset, MM-Diffusion's generation quality significantly deteriorates when using the DPM-Solver method, whereas MGCT can maintain or even surpass DDPM's generation quality when using the DPM-Solver method, further underscoring MGCT's advantage. In conclusion, MGCT excels in fast generation scenarios, especially when fewer inference steps are involved, and it clearly outperforms MM-Diffusion under similar conditions. This balance between efficiency and generation quality offers a promising approach for practical applications.

4.3 Ablation Study

To verify the necessity and effectiveness of the proposed method, we designed two ablation experiments.

Effect of loss function ratio (RQ2). First, to determine the appropriate value of w, we investigated the impact of different w ratios on the optimal re-

Training Method	**Inference Step**	**AIST++**			**Landscape**		
		FVD ↓	**KVD ↓**	**FAD ↓**	**FVD ↓**	**KVD ↓**	**FAD ↓**
MM-Diffusion	DPM-Solver-9step	921.407	190.405	2.495	538.759	22.575	1.794
	DPM-Solver-10step	716.247	153.138	2.513	483.957	17.436	1.618
	DPM-Solver-12step	344.164	61.109	2.402	421.588	13.334	1.858
	DPM-Solver-15step	335.284	68.558	2.25	508.294	17.245	1.855
	DPM-Solver-20step	345.724	66.731	1.604	369.379	9.891	1.256
MGCT (Ours)	DPM-Solver-9step	861.514	191.27	2.384	644.841	33.461	1.685
	DPM-Solver-10step	552.488	109.934	2.333	532.555	22.002	1.77
	DPM-Solver-12step	279.384	36.776	2.305	387.195	14.359	1.874
	DPM-Solver-15step	245.689	35.568	2.41	398.043	12.195	1.3
	DPM-Solver-20step	**209.729**	**20.919**	**1.558**	**313.545**	**7.457**	**1.216**

Table 1: Comparison of the performance of MGCT and MM-Diffusion under different inference steps, using the fast generation method DPM-Solver for sampling.

Dataset	Training Method	FVD↓	KVD↓	FAD↓
AIST++	MM-Diffusion	**198.469**	**27.572**	**1.347**
	MGCT(Ours)	212.669	28.654	1.545
Landscape	MM-Diffusion	353.283	**9.030**	1.521
	MGCT(Ours)	**330.745**	9.777	**1.489**

Table 2: Comparison of the optimal performance of MGCT and MM-Diffusion, using the slow generation method DDPM for sampling.

Dataset	w	FVD↓	KVD↓	FAD↓
AIST++	0.7	**209.729**	**20.919**	1.558
	0.5	335.318	54.875	**1.467**
Landscape	0.7	364.452	9.499	1.225
	0.5	**313.545**	**7.457**	**1.216**

Table 3: Comparison of the optimal generation results of MGCT with different w values across various datasets. Since a large w can lead to convergence difficulties, while a small w may not provide sufficient constraint for adjacent steps, we only consider $w = 0.5, 0.7$ in this experiment.

Dataset	MCT	FVD↓	KVD↓	FAD↓
AIST++	MS	593.72	142.05	1.569
	w/	**209.729**	**20.92**	**1.558**
	w/o	752.826	156.86	2.015
Landscape	MS	509.041	27.35	1.242
	w/	**313.545**	**7.457**	**1.216**
	w/o	653.884	28.41	1.347

Table 4: Comparison of the optimal values for MGCT with and without MCT. For easier comparison, we include MM-Start (MS) as the baseline after applying MCT.

sults for MGCT, trained on the AIST++ and Landscape datasets. We are able to find from Table 3 that when w is 0.7, MGCT achieves the best results on the AIST++ dataset; when w is 0.5, MGCT achieves the best results on the Landscape dataset. This indicates that the value of w has a significant impact on MGCT's performance, and different datasets may require different w values to optimize results. Additionally, compared to Table 1, we observe that regardless of the value of w, MGCT consistently outperforms MM-Diffusion. This demonstrates that the introduction of multimodal consistency constraints significantly enhances the generation quality.

Effect of MCT (RQ3). Next, we will train MGCT from scratch without MCT and compare its best results with those model trained with MCT to verify the necessity of MCT. For a fair comparison, each model will be trained for the same number of iterations, with w set to 0.7 for all models. We are able to find from Table 4 that MGCT trained without MCT shows a significant reduction in generation quality compared to its counterpart trained with MCT, and it even struggles to reach the performance level of MM-Diffusion. This indicates that MCT is useful for our method as it provides low-cost prior knowledge, significantly improving the generation quality. For this phenomenon, we

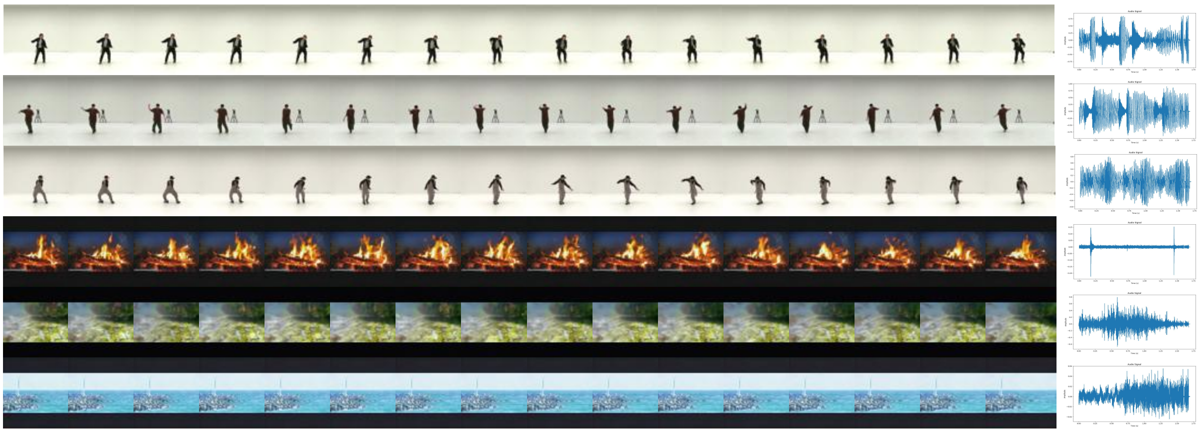

Figure 2: Generated video frames and audio spectrograms from MGCT. The specific content includes three types of street dance as well as three natural landscapes: flames, underwater scenes, and a seaside lighthouse.

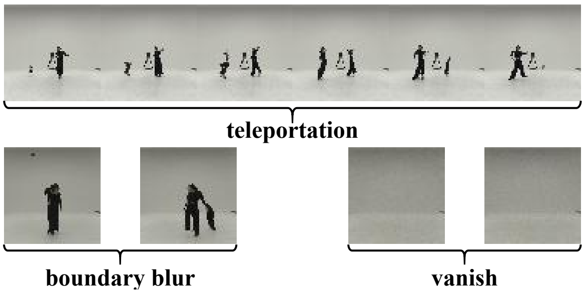

Figure 3: Samples generated by MM-Diffusion exhibit issues such as object teleportation, blurred boundaries, and object disappearance.

believe that if the MGCT model is trained entirely from scratch, it becomes difficult for the model to balance the goal of improving the quality of generated samples while also maintaining consistent denoising capabilities across adjacent steps. Specifically, in the initial stages of the generation process, the denoising ability at each step tends to be relatively weak, but the distances between the steps are small enough to meet the training objectives of consistency models. However, this creates a conflict with the goal of enhancing sample quality, which can result in a sharp drop in training efficiency or even prevent the model from reaching the desired final state.

4.4 Visualization

Generate sample visualisation (RQ4). As shown in Figure 2, the video frames and audio spectrograms generated by MGCT are presented. For the dance motion video, the dynamic performances of the dancer in different poses and movements are displayed. Each frame captures subtle changes in the dancer's movements, showcasing fluidity and rhythm that aligns with the musical beats. In addition, the consistency of the clothing is well maintained throughout the video. For the natural landscape video, elements such as fire, flowing water, seaside, and lighthouse are depicted with rich colors and notable dynamic changes. These scene changes correspond to natural environmental sounds. In Figure 3, it can be observed that when MM-Diffusion uses fast sampling methods such as DPM-Solver, issues such as object teleportation between adjacent frames, unclear boundaries, the appearance of extra limbs, and empty backgrounds are prone to occur. MGCT, however, overcomes these problems effectively. Overall, under the fast sampling method, MGCT successfully achieves high-quality audio-video generation with consistency in both temporal information and content.

5 Conclusion

In this paper, we propose MGCT, a method of Multimodal Generation with Consistency Transferring. Our work utilizes MCT for the initial low-cost learning of data features and employs LCT to constrain the outputs of adjacent steps, thereby enhancing denoising capability and reducing training time costs. The proposed MGCT can generate higher-quality and more stable samples while overcoming the challenges of training with consistency constraints. Through model comparisons and ablation studies, we achieved good performance on widely used datasets, validating the effectiveness and necessity of our proposed innovations.

Limitations

Although MGCT can improves the speed and quality of audio-video generation, we also found that the proposed method has issues such as insufficient sampling speed and excessive denoising during sampling. Going forward, we will optimize the model by addressing aspects such as cross-modal alignment and sampling method adaptation to make audio and video generation more lightweight and stable.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grants No. 62202439). This work was also supported by the advanced computing resources provided by the Supercomputing Center of the USTC.

References

Huixia Ben, Shuo Wang, Meng Wang, and Richang Hong. 2024. Pseudo content hallucination for unpaired image captioning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 320–329.

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

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Dan Guo, Shuo Wang, Qi Tian, and Meng Wang. 2019. Dense temporal convolution network for sign language translation. In *IJCAI*, pages 744–750.

Yanbin Hao, Shuo Wang, Yi Tan, Xiangnan He, Zhenguang Liu, and Meng Wang. 2022. Spatio-temporal collaborative module for efficient action recognition. *IEEE Transactions on Image Processing*, 31:7279–7291.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606.

Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. 2022. Sound-guided semantic video generation. In *European Conference on Computer Vision*, pages 34–50. Springer.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787.

Jinda Lu, Shuo Wang, Xinyu Zhang, Yanbin Hao, and Xiangnan He. 2023. Semantic-based selection, synthesis, and supervision for few-shot learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3569–3578.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. 2023a. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. 2023b. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. 2024. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

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

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. 2023. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022a. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022b. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models. *arXiv preprint arXiv:2303.01469*.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2024. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36.

Shuo Wang, Dan Guo, Xin Xu, Li Zhuo, and Meng Wang. 2019. Cross-modality retrieval by joint correlation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2s):1–16.

Shuo Wang, Dan Guo, Wen gang Zhou, Zheng jun Zha, and Meng Wang. 2018. Connectionist temporal fusion for sign language translation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1483–1491.

Shuo Wang, Jinda Lu, Haiyang Xu, Yanbin Hao, and Xiangnan He. 2024a. Feature mixture on pre-trained model for few-shot learning. *IEEE Transactions on Image Processing*, 33:4104–4115.

Shuo Wang, Xinyu Zhang, Yanbin Hao, Chengbing Wang, and Xiangnan He. 2022. Multi-directional knowledge transfer for few-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3993–4002.

Shuo Wang, Xinyu Zhang, Meng Wang, and Xiangnan He. 2024b. Symmetric hallucination with knowledge transfer for few-shot learning. *IEEE Transactions on Multimedia*.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. 2023. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*.

Jie Xiao, Kai Zhu, Han Zhang, Zhiheng Liu, Yujun Shen, Yu Liu, Xueyang Fu, and Zheng-Jun Zha. 2023. Ccm: Adding conditional controls to text-to-image consistency models. *arXiv preprint arXiv:2312.06971*.

Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. 2023. Moviefactory: Automatic movie creation from text using large generative models for language and images. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9313–9319.

Xingyu Zhu, Shuo Wang, Jinda Lu, Yanbin Hao, Haifeng Liu, and Xiangnan He. 2024. Boosting few-shot learning via attentive feature regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7793–7801.