CMMD: Contrastive Multi-Modal Diffusion for Video-Audio Conditional Modeling

Ruihan Yang[1], Hannes Gamper[2], and Sebastian Braun[2]

[1] University of California Irvine**
[2] Microsoft Research
ruihan.yang@uci.edu
{hannes.gamper,sebastian.braun}@microsoft.com

Abstract. We introduce a multi-modal diffusion model tailored for the bi-directional conditional generation of video and audio. We propose a joint contrastive training loss to improve the synchronization between visual and auditory occurrences. We present experiments on two datasets to evaluate the efficacy of our proposed model. The assessment of generation quality and alignment performance is carried out from various angles, encompassing both objective and subjective metrics. Our findings demonstrate that the proposed model outperforms the baseline in terms of quality and generation speed through introduction of our novel cross-modal easy fusion architectural block. Furthermore, the incorporation of the contrastive loss results in improvements in audio-visual alignment, particularly in the high-correlation video-to-audio generation task.

1 Introduction

Multi-media generation with diffusion models has attracted extensive attention recently. Following breakthroughs in image [22] and audio generation [17], multi-media generation like video remains challenging due to increased data and content size and the added complexity of dealing with both audio and visual components. Challenges for generating multi-modal content include 1) time variant feature maps leading to computationally expensive architecture and 2) audio and video having to be coherent and synchronized in terms of semantics and temporal alignment.

Existing research has predominantly concentrated on unidirectional cross-modal generation, such as producing audio from video cues [19, 35] and vice versa [10, 15]. These approaches typically employ a conditional diffusion model to learn a conditional data distribution $p(x|y)$. Although these models have shown considerable promise, their unidirectional nature is a limitation to learn joint multi-modal representations and general bi-directional use. However, Bayes' theorem elucidates that a joint distribution can be decomposed into $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$, suggesting that the construction of a joint distribution inherently encompasses bi-directional conditional distributions. With the

** Work done while intern at Microsoft Research

advent of the iterative sampling procedure in diffusion models, classifier guidance [2, 8, 28] has emerged as a viable approach for training an unconditional model capable of conditional generation. This approach has been extensively adopted in addressing the inverse problems associated with diffusion models, such as image restoration [12] and text-driven generation [23].

MM-diffusion [25] represents a groundbreaking foray into the simultaneous modeling of video and audio content. The architecture employs a dual U-Net structure, interconnected through cross-attention mechanisms [31], to handle both video and audio signals. Although MM-diffusion demonstrates impressive results in terms of *unconditional* generation quality, it has two major limitations: Firstly, it's random-shift cross-attention mechanism is still complex and it relies on a super-resolution up-scaling model to improve image quality. Secondly, the focus has been on unconditional generation, while we focus on conditional generation and improve the evaluation methodology.

In this study, we introduce an improved multi-modal diffusion architecture with focus on bi-directional *conditional* generation of video and audio. This model incorporates an optimized design that more effectively integrates video and audio data for conditional generation tasks. More importantly, we leverage a novel joint contrastive diffusion loss to improve alignment between video and audio pairs. Our experiments on two different dataset employ both subjective and objective evaluation criteria. We achieve superior quality than the baseline and stronger synchronization.

The key contributions can be summarized as follows:

- We present an optimized version of the multi-modal *latent-spectrogram* diffusion model, featuring a pretrained video autoencoder, a vocoder and an easy fusion mechanism. This design aims to more effectively integrate cross-modality information between video and audio, while also enhancing *conditional* sampling quality.
- Drawing inspiration from uni-modal contrastive learning, we propose a novel contrastive loss function tailored for the joint model. This function is instrumental in enhancing the alignment accuracy for the conditional generation of video-audio pairs.
- Our experimental evaluations, performed on two distinct datasets, AIST++ [16] and EPIC-Sound [9]. We propose to use metrics with improved correlation human perception and practical relevance compared to prior work in the field. The assessments, based on a range of subjective and objective metrics demonstrate that our method outperforms the existing MM-diffusion [25] in terms of quality, as well as non-contrastive variants in terms of temporal synchronization.

2 Method

In this section, we provide an overview of the diffusion model employed, followed by a description of the intricacies of the architecture design of the proposed model. Finally, we introduce the joint contrastive loss that enhances the

alignment of video and audio components. An overview of our model is shown in Fig. 1.

2.1 Video-Audio Joint Diffusion Model

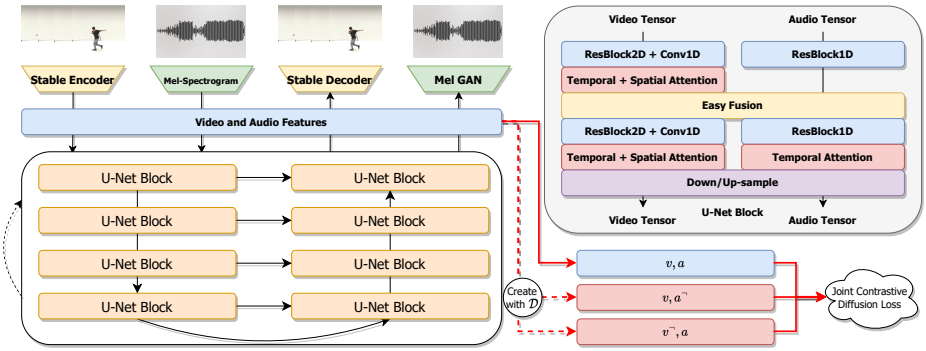

Fig. 1: Overview of our proposed architecture and method. The detailed implementation of each U-Net block is depicted in the upper right corner and the intuition of our design choice of easy fusion is available in Appendix. Training of the diffusion model is performed on latent-spectrogram space.

Denoising diffusion models introduced a practical objective function for training the reverse process [6, 7, 26]:

$$\mathbb{E}_{n,\epsilon}[||\epsilon - \epsilon_\theta(\mathbf{x}_n(\mathbf{x}_0), n)||^2] \text{ or } \mathbb{E}_{n,\epsilon}[||\mathbf{v} - \mathbf{v}_\theta(\mathbf{x}_n(\mathbf{x}_0), n)||^2] \tag{1}$$

where ϵ_θ represents the most commonly used parameterization in previous works [2, 7, 11, 24, 25, 27], and $\mathbf{v}_\theta$ (velocity) has also shown promising results with a more stable training process [26]. We adopt the latter method to train our model.

Video-Audio Modeling Our approach to video-audio joint modeling follows a design analogous to the uni-modal diffusion model. Here, the data point $\mathbf{x}$ comprises two modalities: the video signal $v_{0..N}$ and audio signal $a_{0..N}$. Consequently, the optimization objective resembles the form in Eq.1:

$$\mathcal{L}_{\text{jdiff}} = \mathbb{E}_{n,\epsilon}[||\mathbf{v} - \mathbf{v}_\theta(v_n, a_n, n)||^2] \tag{2}$$

where $\mathbf{v}$ represents the velocity parameterization for both video and audio, specifically $\mathbf{v} = [\sqrt{\alpha_n}\epsilon_v - \sqrt{1-\alpha_n}v_0, \sqrt{\alpha_n}\epsilon_a - \sqrt{1-\alpha_n}a_0]$. This implies that the model $\mathbf{v}_\theta$ simultaneously predicts two outputs, embodying a joint diffusion model that effectively manages both modalities.

Guided Conditional Generation An intriguing aspect of diffusion models is their capacity to enable conditional generation through guidance from a classifier, even in the context of models trained without conditioning [2]. Typically, this guidance method involves an additional classifier, $p_\phi(y|x)$, and utilizes the gradient term $\nabla_x p_\phi(y|x)$ to adjust the sampling direction during the denoising process.

However, in our model, which considers both video and audio modalities, we can employ a more straightforward *reconstruction guidance* approach [8]. For the video-to-audio generation case, we can formalize conditional generation as follows (audio-to-video shares a similar formulation).

$$\begin{aligned} \hat{a}_0 &= \bar{a}_0 - \underbrace{\lambda\sqrt{\alpha_n}\nabla_{a_n}||v_0 - \hat{v}_0||^2}_{\text{reconstruction guidance}} \\ a_{n-1} &= \sqrt{\alpha_{n-1}}\hat{a}_0 + \sqrt{1-\alpha_{n-1}}\epsilon_a \end{aligned} \tag{3}$$

where the gradient guidance is weighted by λ and $\bar{a}_0$ is the unguided noisy audio signal reconstruction at denoising step n. In the case of $\lambda = 0$, the generation scheme is equivalent to the *replacement* method [28]. Both ϵ_a and ϵ_v are drawn from an isotropic Gaussian prior at the start of the iteration. Therefore, these equations depict an intermediate stage of the conditional generation process using the DDIM sampling method [27]. Although the speed of sampling is not the primary focus of our model, alternative ODE or SDE solvers can be employed to expedite the denoising sampling process [11, 18].

2.2 Joint Contrastive Training

To improve the synchronization of video and audio in our model, we utilize principles of contrastive learning [21]. This approach has proven effective in maximizing the mutual information $I(a; v)$ for video-to-audio conditional generation [19,35]. The CDCD [35] method seamlessly integrates contrastive loss into the video-to-audio conditional diffusion models, as given by

$$\begin{aligned} \mathcal{L}_{\text{cont}} &:= \mathbb{E}_A \log\left[1 + \frac{p_\theta(a_{0:N})}{q(a_{0:N}|v_0)}\, M\mathbb{E}_{A'}\left[\frac{p_\theta(a^{\neg}_{0:N}|v_0)}{q(a^{\neg}_{0:N})}\right]\right] \\ &\approx \mathcal{L}_{\text{cdiff}}(a_{0:N}, v_0) - \eta \sum_{a^{\neg}_0 \in A'} \mathcal{L}_{\text{cdiff}}(a^{\neg}_{0:N}, v_0) \end{aligned} \tag{4}$$

where the set A includes the correct corresponding audio samples, while A' contains the mismatched negative samples of A. L_{cdiff} denotes the unimodal conditional diffusion loss, with v representing the conditioning videos and M indicating the number of negative samples. To streamline the training process, we replace M with a weighting term η, eliminating the need to generate M negative samples at each training step. This means at each training step, we can sub-sample a batch of $a^{\neg}$ from the M samples for computational efficiency.

The above formulation pertains to training a classifier-free conditional diffusion model. To adapt this approach to our joint diffusion loss, as described in Eq.2, we observe that we are training an implicit conditional diffusion model

$p_\theta(a_{n-1}|a_n, v_n)$. Eq.3 demonstrates that v_n can be directly calculated during conditional generation:

$$v_n \sim q(v_n|v_0) = \sqrt{\alpha_n}v_0 + \sqrt{1-\alpha_n}\epsilon_v \tag{5}$$

which implies that $v_{1:N}$ is fixed with a given ϵ_v and v_0. Given this relationship between v_n and v_0, we have following approximation $p_\theta(a_{n-1}|a_n, v_0) \approx p_\theta(a_{n-1}|a_n, v_n)q(v_n|v_0)$. Thus, we can bridge Eq.4 to our jointly trained multimodal diffusion model. For audio-to-video generation, we can follow the same method above by swapping v and a. Finally, the resulting joint contrastive loss can be represented by the following three terms:

$$\begin{aligned}\mathcal{L}_{\text{cont}} = \mathcal{L}_{\text{jdiff}}(a_{0:N}, v_{0:N}) &- \eta\mathbb{E}_{a_0^\neg \sim A'}\mathcal{L}_{\text{jdiff}}(a_{0:N}^\neg, v_{0:N}) \\ &- \eta\mathbb{E}_{v_0^\neg \sim V'}\mathcal{L}_{\text{jdiff}}(a_{0:N}, v_{0:N}^\neg)\end{aligned} \tag{6}$$

where V' denotes the set of negative samples for a_0 and η adjusts the weight of the contrastive term. It's important to note that, instead of iterating over all the V' and A' samples, we choose to randomly draw a subset from them per gradient descent step to reduce GPU memory consumption.

Creating Negative Samples In absence of a pre-existing high-quality dataset for contrastive learning, we can generate negative samples through data augmentation. Specifically, we employ the following methods to create V' and A' in the context of paired positive data a, v. For brevity, we will only outline the generation of negative audio samples $a^\neg$. The creation of negative videos $v^\neg$ follows a similar formulation:

- *Random Temporal Shifts*: We apply random temporal shifts to a, moving the signal backward or forward by a random duration within some hundreds of milliseconds.
- *Random Segmentation and Swapping*: We randomly draw a separate audio segment, denoted as a_d, with the same length as a. Subsequently, we sample a random split point on both a_d and a, allowing us to construct $a^\neg$ as either concatenate$(a_d^{\text{left}}, a^{\text{right}})$ or concatenate$(a^{\text{left}}, a_d^{\text{right}})$.
- *Random Swapping*: In this method, we randomly select a different audio segment, a_d, of the same length as a, and substitute a with a_d.

The detailed training procedure is outlined in Appendix Algorithm.

3 Experiments

Datasets Our evaluation leverages two datasets, each offering unique challenges and scenarios within the audio-video domain: **AIST++** [16] is derived from the AIST Dance Database [29]. This dataset features street dance videos with accompanying music. It serves a dual purpose in our evaluation, being used for both video-to-audio and audio-to-video tasks. The **EPIC-Sound** [9] dataset consists

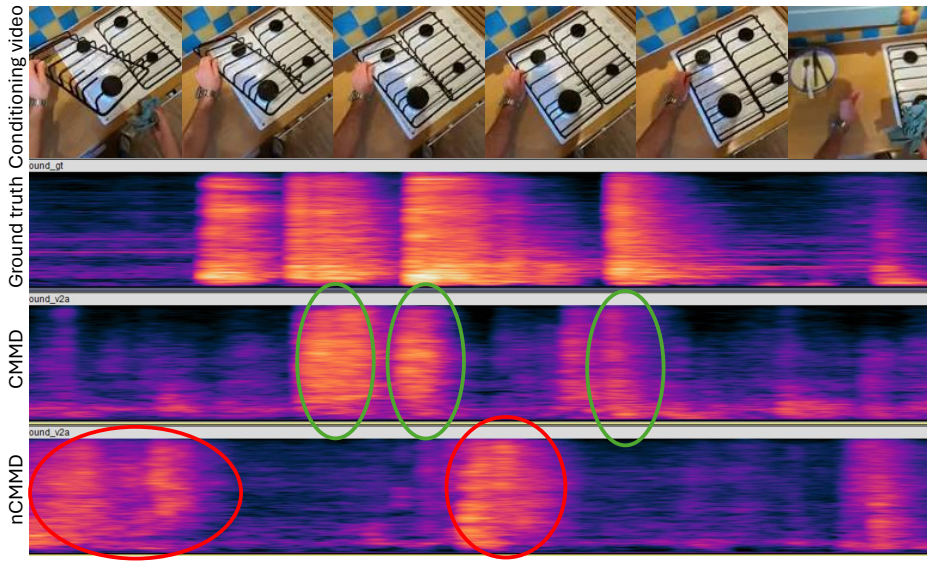

Fig. 2: Conditioning video (top) with ground truth spectrogram below. The two bottom spectrograms show the generated audio with CMMD and nCMMD conditioned on the video. Sound events are highlighted with a green circle for matches and a red circle for mismatches.

of first-person view video recordings that capture a variety of kitchen activities, such as cooking, that are characterized by a strong audio-visual correlation. Due to the significant motion and camera movement in the videos, which complicates visual learning, we use EPIC-Sound exclusively for video-to-audio evaluation.

Baselines The MM-Diffusion model [25] stands as the only known baseline capable of handling both video-to-audio and audio-to-video synthesis tasks. For our comparison, we employed the official MM-Diffusion implementation, utilizing weights trained on the 1.6 s 10fps AIST++ dataset at a resolution of 64×64. Additionally, we present results from nCMMD, a variant of our CMMD model that does not incorporate contrastive loss.

Feature Extraction & Data Preprocessing We sampled 18 frames from 10 fps video sequences and the corresponding 1.8s audio at 16kHz. Video frames underwent center cropping and resizing to a 128×128 resolution, or optionally downsampling to 64×64 for a comparison with the MM-Diffusion baseline. The audio samples represented in a Mel Spectrogram have 80 channels and 112 time steps. During test time, we use twice the training sequence length, i.e., 36 video frames, if not specified otherwise.

As outlined in Appendix, we encoded videos using the Gaussian VAE from the Stable Diffusion project [24], which effectively reduces image resolution by a fac-

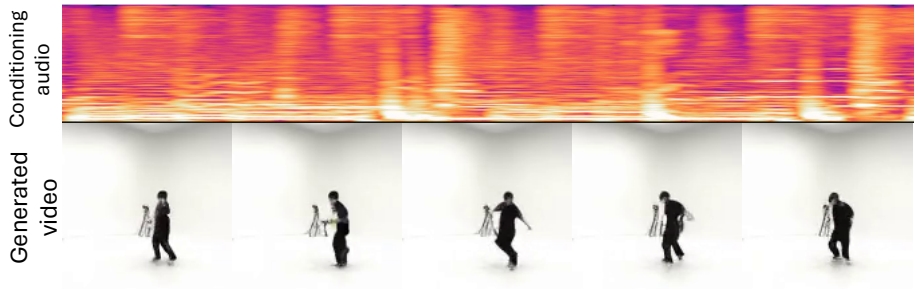

Fig. 3: Generated video with CMMD conditioned on the audio spectrogram.

tor of eight in both width and height. We utilized the pre-trained model weights[3] without further fine-tuning. For audio features, we transformed waveforms sampled at 16 kHz into 80-bin mel spectrograms using a Short-Time Fourier Transform (STFT) with a 32 ms window and 50% overlap, yielding a time resolution of 16 ms. The MelGAN vocoder was improved by the loss weightings from HifiGAN [14] and notably improved by training on sequences of 4 s, as opposed to the originally suggested 0.5 s. This adjustment aligns with the MelGAN architecture's receptive field of approximately 1.6 s. The vocoder was trained on the entire AudioSet [4] to ensure a broad sound reconstruction capability.

3.1 Metrics

Fréchet Distance Objective metrics to capture the perceived quality of video and audio are often difficult to develop and have many imperfections. Especially in generative tasks, where new content is created and no ground truth is available, such metrics are to be used with care. Popular approaches are statistical metrics, which compare generated and reference distributions in some embedding space, such as the *Fréchet Audio Distance (FAD)* [13] and *Fréchet Video Distance (FVD)* [30]. We assess FVD in a pairwise manner [32, 33]: calculating the score between the 5 times conditional generation results and the corresponding ground truth test sets. To measure audio quality, we calculate FAD using CLAP embeddings [3], which have been shown recently in [5] to represent acoustic quality much better than the widely used VGGish features. FAD scores are calculated using the FAD toolkit [5] both individually for each generated sample and for the entire set of samples generated by one model, using the test set as a reference. Additionally, we also consider KVD [1] as a complementary metric of visual quality for video contents.

Temporal Alignment For the dancing videos from AIST++, to evaluate the temporal alignment of generated music, we use a beat tracking approach similarly as in [35] to measure the rhythmic synchronicity. The music beats are estimated

[3] https://huggingface.co/stabilityai/sd-vae-ft-mse

Models	CMMD		nCMMD		MM-Diff	
	FVD	KVD	FVD	KVD	FVD	KVD
16 frames (64)	**611**	58	703	83	726	**48**
18 frames (64)	**749**	**70**	799	187	757	71
32 frames (64)	765	53	**708**	**47**	871	68
18 frames (128)	**934**	78	1036	136	N/A	
36 frames (128)	973	**49**	**882**	49	N/A	

Table 1: FVD and KVD results for different frame settings on AIST++ dataset. Numbers in parentheses indicate the resolution of the evaluated frames.

using librosa [20] beat tracker and the hit rate between beats of generated and ground truth audio is computed. We propose to use a tolerance of ±100 ms, which corresponds approximately to the average perceivable audio-visual synchronicity thresholds found in literature [34]. For reference, we also show results using a larger tolerance of ±500 ms,which is equivalent to the 1 s quantization used in [35]. While this significantly improve accuracy numbers, we consider this a very inaccurate, close to random metric, as a 1 s window already contains 2 beats at a average song tempo of 120 beats per minute. Since the beat tracking method is applicable only to musical content, we reserve the alignment assessment for EPIC-Sound to subjective evaluation.

Subjective Evaluation We conducted a user study with 14 participants to evaluate the audio-visual quality and synchronicity. Participants were recruited lab internally on voluntary basis without restrictions except unimpaired vision and hearing. No further demographical information was collected for privacy reasons. For each example, we asked two or three questions about the quality of the generated content and the temporal alignment of video and audio events on MOS scales from 1 (worst) to 5 (best). Specifically, for *generated video*, we asked to rate the *video quality* and the *temporal alignment*. For *audio generation* from AIST++ dance videos, we asked to rate separately the *acoustic* and *musical quality*, and the *temporal synchronization* of the dancer to the music. For the EPIC-Sound cases, we asked to rate the *acoustic* and *semantic quality*, and the *temporal synchronization* of events. Semantic quality refers to whether the type of sounds heard make sense given the scene seen in the video without paying attention to temporal synchronization.

3.2 Objective Evaluation Results

The results for Fréchet Video Distance (FVD) and Kernel Video Distance (KVD) comparing the proposed model and baseline models are detailed in Table 1. The findings reveal that (n)CMMD consistently outperforms MM-Diffusion across a variety of resolutions and sequence lengths. Specifically, CMMD demonstrates a marginal superiority over nCMMD in shorter sequences. Conversely, nCMMD

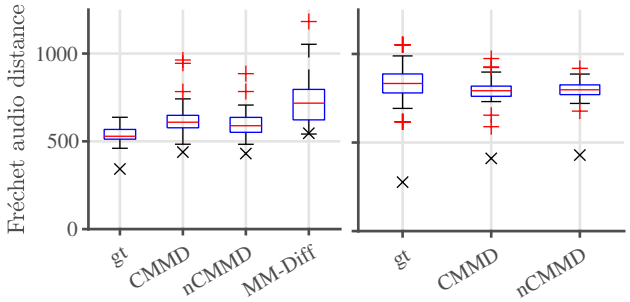

Fig. 4: Per-sample (boxes) and per-set (×) Frechet audio distance (FAD) results for AIST++ (left) and EPIC-Sound (right). FAD is calculated for 50 output samples of each model using CLAP embeddings with the respective test set as reference. Boxes show the per-sample FAD distribution of these 50 samples, with red markers indicating outliers beyond the whiskers which extend to 1.5 times the interquartile range. Note that the per-set FAD scores for ground truth (gt) are larger than zero as only the small subset of the test set used in the evaluation is compared to the whole test set used as reference. Comparing FAD scores for identical set sizes avoids sample size bias [5].

exhibits slightly better quality in longer sequences, aligning with our subjective assessments. MM-Diffusion, however, performs better only in terms of KVD for low-resolution, short video sequences, which is the specific condition under which this model was trained.

Fig. 4 illustrates the comparison of audio quality in a video-to-audio generation scenario. Our CMMD model surpasses the baseline in AIST++ music audio quality, in terms of both per-sample FAD [5] and batch FAD metrics. There is no significant difference between CMMD and nCMMD for both datasets.

Table 2 presents the beat alignment results for the AIST++ audios. The table compares three different methods: CMMD, nCMMD, and MM-Diffusion. In terms of beat tracking accuracy within a 100 ms tolerance, CMMD performs the best, showing a improvement of 1-4%. As mentioned in 3.1, we do not consider the results for tolerance of 500 ms meaningful as it allows very coarse and ambiguous beat matches, so we do not suggest to draw conclusions from this setting. It is only notable that this large inaccurate tolerance doubles accuracy numbers, which we find misleading.

In the generation processes of audios for EPIC-Sound and videos for AIST++, our primary reliance is on subjective evaluation, given the absence of robust metrics. To supplement this assessment, we present EPIC-Sound audio generation visualization provided in Fig. 2, where we can observe that CMMD has better alignment with the ground truth than nCMMD in terms of temporal sound event alignment. Additionally, Fig. 3 presents a qualitative sample showcasing audio-to-video generation in the context of AIST++.

Hitrate tolerance	CMMD	nCMMD	MM-Diff (Ruan2023)	comments
±500 ms	89%	91%	89%	not suggested tolerance
±100 ms	**45%**	44%	41%	

Table 2: Comparison of Beat Tracking Accuracy (AIST++). The values in parentheses indicate the allowable margin of error for beat timing, with a smaller window representing a stricter standard. Higher hit rates within lower tolerance thresholds signify superior temporal alignment.

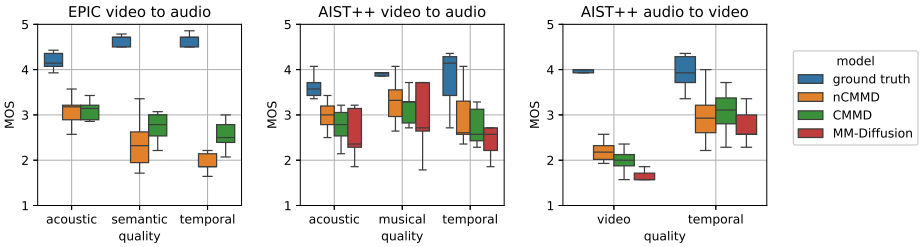

Fig. 5: Subjective results from user study for EPIC-Sound video conditioned audio generation (left), AIST++ dance video conditioned audio generation (center), and audio conditioned video generation (right).

3.3 Subjective Evaluation Results

In the subjective evaluation we used 85 videos. We used 5 different conditions (audio or video as conditioning), two different sample generations per CMMD and nCMMD model, one sample each per ground truth and baseline. For AIST++ we evaluated audio to video and video to audio generation. For EPIC-Sound, we evaluated only video to audio and there is no MM-Diffusion baseline available.

The Mean Opinion Scores (MOS) are shown as boxplots in Fig. 5, where the black bars show then median, the boxes show the inter-quartile range, and the whiskers show the minimum and maximum values. Additionally, we test statistical significance using the Wilcoxon signed-rank test with p-value < 0.05 to analyze close cases. We can see that the raters reliably detected the ground truth samples attributing it the highest score, although often the scale was not used fully. For the generated dance visuals from AIST++ audio (Fig. 5 right), we can observe a significantly higher rating of our proposed models over MM-Diffusion baseline. The nCMMD model has a slightly higher video quality with p=0.005. The CMMD model shows a trending but non-significant better temporal alignment than nCMMD with $p = 0.327$. CMMD temporal alignment is significantly better than MM-Diffusion baseline with $p = 0.038$.

For audio generation conditioned on AIST++ dance videos (Fig. 5 center), we observe the temporal alignment of CMMD and nCMMD as well as their acoustic quality better than MM-Diffusion. While these differences are smaller, they are statistically significant. nCMMD has slightly better acoustic quality

than CMMD, while there is no significance between CMMD and nCMMD temporal alignment with $p = 0.09$, as can also be seen on their very close medians. nCMMD outperforms MM-Diffusion in musical quality, while the spread is too large to draw conclusions between CMMD and MM-Diffusion on musical quality.

For audio generation conditioned on EPIC-Sound videos (Fig. 5 left), CMMD outperforms nCMMD in terms of semantic quality and temporal alignment due to the use of the contrastive loss, while the acoustic quality is on par.

3.4 Discussion

The results in Fig. 5 on EPIC-Sound video to audio task show a clear benefit of the contrastive loss to enforce stronger both temporal synchronization and semantic alignment without sacrificing audio acoustic quality. In AIST++, the contrastive loss improves the temporal synchronization MOS for the audio to video condition, while it is inconclusive for the video to audio condition. However for this condition, the $100ms$ beat tracking metric in Tab. 2 still indicates a minor synchronization improvement. Interestingly, on AIST++ it seems that the model trades off a small amount of quality in favor of better synchronization, while objective metrics like FVD, KVD and FAD are on par or fluctuating depending on condition. In general, the temporal synchronization results are less pronounced for the AIST++ dance data, possibly due to the fact that the alignment of human dancers with music may be harder to judge for several reasons: 1) the dancers may vary in tempo or their internal rhythm may be judged in ambiguous ways. 2) being off by one or two full beats may appear as being in sync again.

References

1. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
2. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
3. Elizalde, B., Deshmukh, S., Ismail, M.A., Wang, H.: Clap learning audio concepts from natural language supervision. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095889
4. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proc. IEEE ICASSP 2017. New Orleans, LA (2017)
5. Gui, A., Gamper, H., Braun, S., Emmanouilidou, D.: Adapting frechet audio distance for generative music evaluation. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2024)
6. Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., Guo, B.: Efficient diffusion training via min-snr weighting strategy. arXiv preprint arXiv:2303.09556 (2023)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)

8. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 8633–8646. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf
9. Huh, J., Chalk, J., Kazakos, E., Damen, D., Zisserman, A.: EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. In: IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP) (2023)
10. Jeong, Y., Ryoo, W., Lee, S., Seo, D., Byeon, W., Kim, S., Kim, J.: The power of sound (tpos): Audio reactive video generation with stable diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7822–7832 (2023)
11. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022)
12. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Advances in Neural Information Processing Systems **35**, 23593–23606 (2022)
13. Kilgour, K., Zuluaga, M., Roblek, D., Sharifi, M.: Fr\'echet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms (Jan 2019). https://doi.org/10.48550/arXiv.1812.08466, http://arxiv.org/abs/1812.08466, arXiv:1812.08466 [cs, eess]
14. Kong, J., Kim, J., Bae, J.: HiFi-GAN: Generative adversarial networks forefficient and high fidelity speech synthesis. In: Proceesings of 34th Conference on Neural Information Processing Systems (2020)
15. Lee, S.H., Kim, S., Yoo, I., Yang, F., Cho, D., Kim, Y., Chang, H., Kim, J., Kim, S.: Soundini: Sound-guided diffusion for natural video editing. arXiv preprint arXiv:2304.06818 (2023)
16. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Learn to dance with aist++: Music conditioned 3d dance generation (2021)
17. Liu, H., Tian, Q., Yuan, Y., Liu, X., Mei, X., Kong, Q., Wang, Y., Wang, W., Wang, Y., Plumbley, M.D.: Audioldm 2: Learning holistic audio generation with self-supervised pretraining (2023)
18. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems **35**, 5775–5787 (2022)
19. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. arXiv preprint arXiv:2306.17203 (2023)
20. McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., van Niekirk, B., Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., Halvachs, E., Thomé, C., Robert-Stöter, F., Bittner, R., Wei, Z., Weiss, A., Battenberg, E., Choi, K., Yamamoto, R., Carr, C., Metsai, A., Sullivan, S., Friesch, P., Krishnakumar, A., Hidaka, S., Kowalik, S., Keller, F., Mazur, D., Chabot-Leclerc, A., Hawthorne, C., Ramaprasad, C., Keum, M., Gomez, J., Monroe, W., Morozov, V.A., Eliasi, K., nullmightybofo, Biberstein, P., Sergin, N.D., Hennequin, R., Naktinis, R., beantowel, Kim, T., Åsen, J.P., Lim, J., Malins, A., Hereñú, D., van der Struijk, S., Nickel, L., Wu, J., Wang, Z., Gates, T., Vollrath, M., Sarroff, A., Xiao-Ming, Porter, A., Kranzler, S., Voodoohop, Gangi, M.D., Jinoz, H., Guerrero, C., Mazhar, A., toddrme2178, Baratz, Z., Kostin, A., Zhuang, X., Lo, C.T., Campr,

P., Semeniuc, E., Biswal, M., Moura, S., Brossier, P., Lee, H., Pimenta, W.: librosa/librosa: 0.10.1 (Aug 2023). https://doi.org/10.5281/zenodo.8252662, https://doi.org/10.5281/zenodo.8252662
21. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
22. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
23. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
25. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: 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. pp. 10219–10228 (2023)
26. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022)
27. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
28. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
29. Tsuchida, S., Fukayama, S., Hamasaki, M., Goto, M.: Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019. Delft, Netherlands (Nov 2019)
30. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
32. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Advances in Neural Information Processing Systems **35**, 23371–23385 (2022)
33. Yang, R., Srivastava, P., Mandt, S.: Diffusion probabilistic modeling for video generation. arXiv preprint arXiv:2203.09481 (2022)
34. Younkin, A.C., Corriveau, P.J.: Determining the amount of audio-video synchronization errors perceptible to the average end-user. IEEE Transactions on Broadcasting **54**(3), 623–627 (2008). https://doi.org/10.1109/TBC.2008.2002102
35. Zhu, Y., Wu, Y., Olszewski, K., Ren, J., Tulyakov, S., Yan, Y.: Discrete contrastive diffusion for cross-modal music and image generation. In: The Eleventh International Conference on Learning Representations (2023)