KeyVID: Keyframe-Aware Video Diffusion for Audio-Synchronized Visual Animation

Anonymous ICCV submission

Paper ID 8

Abstract

Generating video from conditions, such as text, image, and audio, enables both spatial and temporal control, leading to high-quality generation results. Videos with dramatic motions often require a higher frame rate to ensure smooth motion. Currently, most audio-to-visual animation models use uniformly sampled frames from video clips. However, these uniformly sampled frames fail to capture significant key moments in dramatic motions at low frame rates and require significantly more memory when increasing the number of frames directly. In this paper, we propose KeyVID, a keyframe-aware audio-to-visual animation framework that significantly improves the generation quality for key moments in audio signals. Given an image and an audio input, we first localize keyframe time steps from the audio. Then, we use a keyframe generator to generate the corresponding visual keyframes. Finally, we generate all intermediate frames using the motion interpolator. Through extensive experiments, we demonstrate that KeyVID significantly improves audio-video synchronization and video quality across multiple datasets, particularly for highly dynamic motions. The code and demo will be released after acceptance.

1. Introduction

Recent years have witnessed remarkable progress in video generation, driven by advancements in diffusion-based models [1, 3, 24]. These frameworks typically condition the generation process on text prompts and/or image inputs, where the text provides semantic guidance, while the image specifies spatial composition. Despite their success, these methods largely focus on aligning visual outputs with static text or image, leaving dynamic, time-sensitive modalities such as audio underexplored.

Audio-Synchronized Visual Animation (ASVA) [26] aims to animate objects in a static image into a video with motion synchronized with the input audio. To achieve precise synchronization, it is crucial to align key visual actions with their corresponding audio signals. For example, given hammering sounds, the hammer should strike at the exact moment the impact sound occurs. However, this synchronization is constrained by frame rates—AVSyncD [26] operates at 6 FPS, while audio carries fine-grained temporal information, causing key moments to be lost in sparse low frame rate videos (see Fig. 1).

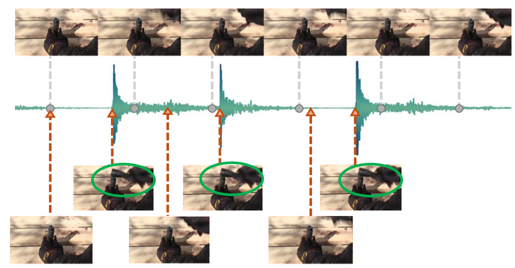


Figure 1. *Top*: Uniformly sampled sparse frames fail to capture the key moments in audio. *Bottom*: Key frames precisely aligned with the hammer strike, matching the critical moments.

Directly training on high frame rate videos incurs substantial computational costs in GPU memory and training time. A common solution adopts a two-stage strategy that generates low frame rate videos then applies frame interpolation [1, 18]. However, this approach struggles in highly dynamic sequences, where critical events may be lost due to the sparsity of initial uniform frames. To ensure accurate audio-visual synchronization while maintaining computation efficiency, we propose **KeyVID**, a **Key**frame-aware **VI**deo **D**iffusion framework. We first develop a keyframe selection strategy that identifies critical moments based on optical flow-based motion score. A *Keyframe Localizer* predicts keyframe positions directly from audio. Instead of uniform downsampling, we train a *Keyframe Generator* that explicitly captures crucial moments without requiring excessive frames. A specialized *Motion Interpolator* then synthesizes intermediate frames between the uneven keyframes. This approach mimics animation workflows in the animation industry where a "Key Animator" establishes crucial moments and a "In-betweener" fills gaps.

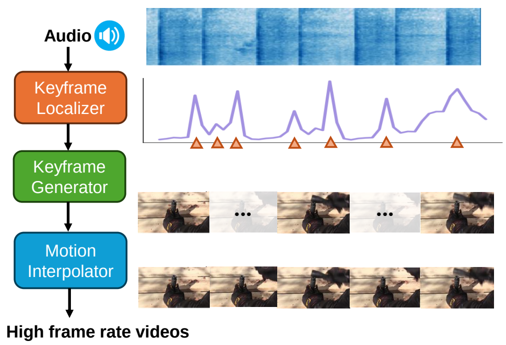


Figure 2. KeyVID video generation pipeline.

We demonstrate that our keyframe-aware approach outperforms state-of-the-art methods in video quality and audio-video synchronization in ASVA task. The main contributions are:

- A keyframe-aware framework that localizes keyframes from audio and generates them as video diffusion model.
- A motion interpolation network enabling high frame rate generation while maintaining efficiency.
- Superior performance in audio-synchronized video generation, particularly for dynamic scenes.

2. Related Work

Video Diffusion Models. Recent video diffusion models [1, 3, 8, 24] generate high-quality videos by learning to denoise Gaussian noise through a reversed diffusion process. For efficiency, latent video generation [2, 24] encodes video into latent space. These models incorporate text [3, 9] and image [6, 24] to guide generation, but typically use uniform frame sampling, limiting their ability under critical motion moments at low frame rates.

Audio-to-Video Generation. While early audio-conditioned video synthesis focused on domain-specific tasks [13, 19], recent work leverage pre-trained audio encoders [4, 5] for general video generation. Some approaches use audio as global features [10, 21], while others [11, 12, 17] consider temporal alignment. AVSyncD [26] introduced time-dependent audio features for finer temporal control, but remains limited by low frame rates (6 FPS) for dynamic motions. Directly increasing frame density requires prohibitive computational resources.

3. Methods

We present KeyVID, a keyframe-aware video generation framework. Given input audio and first frame, we follow a three-stage process to get the the video output (see Fig. 2): (1) the *Keyframe Localizer* first predicts keyframe locations from audio; (2) the *Keyframe Generator* produces the keyframes conditioned on audio and image for each positions; (3) *Motion Interpolator* synthesize all the intermediate frames to obtains a smooth high frame rate videos.

3.1. Keyframe Localization from Audio

We train a *Keyframe Localizer* to infer keyframe locations from audio by exploiting correlations between acoustic events and motion changes. These motion changes are defined by analyzing the motion score from training videos and serve as pseudo labels. Using RAFT [22], we compute optical flow $\mathbf{OF}_t$ between consecutive frames and calculate the motion score as: $M(t) = \sum_{i,j}(|u_t(i,j)| + |v_t(i,j)|)$, where u_t, v_t are horizontal/vertical flow components. The keyframe localizer takes audio spectrograms as input, which consists of pretrained ImageBind [5] as feature extractor and fully connected layers to predict motion scores. The training process is guided by L2 loss.

3.2. Audio-conditioned Keyframe Generation

Our keyframe generator produces T_K keyframes from a T-frame video, conditioned on audio, first frame, and text.

3.2.1. Keyframe Data Selection

Rather than the uniformly sampled T_K frames [24] to train the video diffusion model, we select $T_K \ll T$ keyframes from the peaks and valleys of motion score which represents the most crucial moments of motions in a video clip. We choose the first frame, randomly select up to $\frac{T_K}{2} - 1$ peaks, add valleys between consecutive peaks, then evenly sample remaining frames. This ensures coverage of critical moments while approximating uniform sampling for smooth sequences. The details of the selection algorithm are in Appendix B. The selected keyframe indices $\{t_1, \ldots, t_{T_K}\}$ serve as additional conditions to the following step.

3.2.2. Keyframe Generator Diffusion Model

The keyframe generator introduces two key enhancements: (1)A **Frame index embedding** encodes each frame's absolute position, ensuring coherence when generating non-uniformly distributed frames; (2) **Multi-modal condition features** consist of global text features, and audio and image features extracted from corresponding keyframe timesteps.

As shown in Fig. 3(b), we build upon latent diffusion models [1, 2] with pretrained encoder and decoder from Xing et al. [24]. The latent features are represented as $\mathbf{z} \in \mathbb{R}^{B \times T_k \times C \times h \times w}$ where B denotes the batch size, T_k the number of frames at each denoising step, h and w the spatial dimensions, and C is the feature channels.

Frame Index Embedding. We introduce a embedding layer to encode the absolute index of each keyframe within the original video sequence $\{i_t\}_{t=1}^{T_K}$. The frame index embedding $\mathbf{f_{emb}} \in \mathbb{R}^{B \times T_K \times C}$ is added up with the latent features $\mathbf{z}$ before passing them into the denoising U-Net, ensuring explicit positional information is provided to the network to enable generation of non-uniformly spaced frames.

Audio Feature Conditions. We extract audio features using pretrained ImageBind [5], which encodes spectrograms into global and local tokens capturing semantic and tempo-

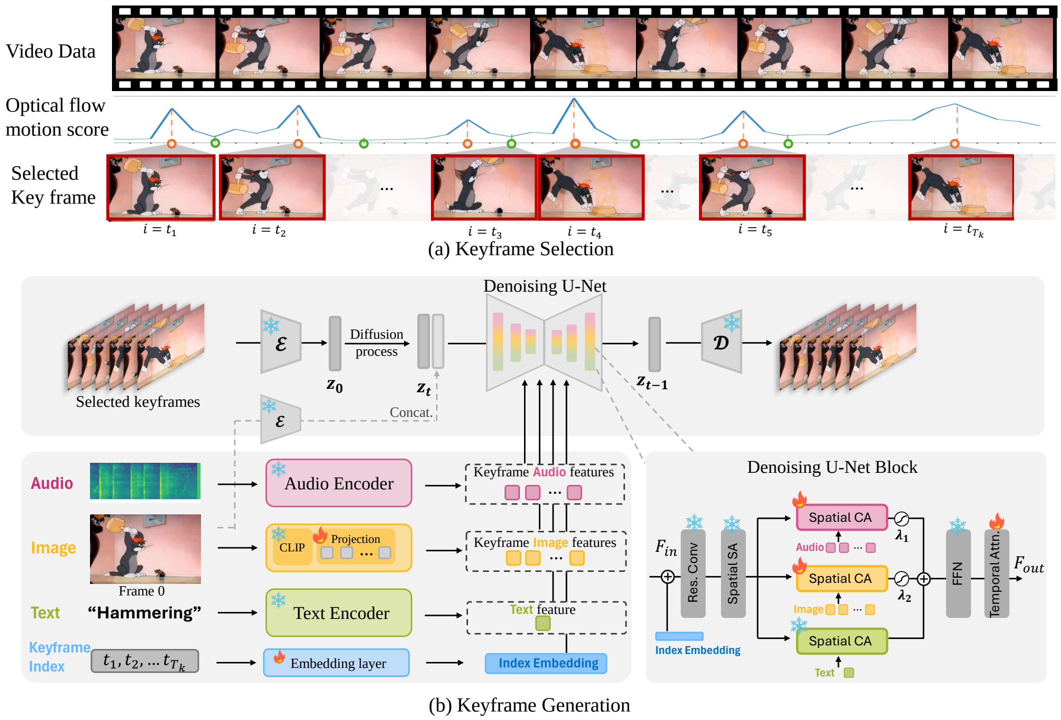


Figure 3. (a) Keyframe selection from motion score curve. (b) Diffusion model for keyframe generation with audio, image, text conditions and frame index embedding.

ral information. We segment the features into T timesteps matching the final video length, then select T_K features at keyframe indices $\{i_t\}_{t=1}^{T_K}$ for cross-attention fusion in the U-Net, ensuring audio-visual synchronization.

Image/Text Feature Conditions. We extract first-frame features using CLIP [16] and project them to T frames via learned tokens, yielding $\mathbf{f}_{\text{img}} \in \mathbb{R}^{B\times T\times C\times H\times W}$. We select T_K features at keyframe indices $\{i_t\}_{t=1}^{T_K}$ for the cross attention during denoising. The text descriptions are encoded using CLIP [16] and repeated for all T_K keyframes.

Feature Fusion. Each conditioning feature (audio, image, and text) is processed separately through spatial cross-attention layers during the denoising step. Given input latent features $\mathbf{F}_{\text{in}}$, one denoising step computes:

$$\begin{aligned}\mathbf{F}_{\text{out}} = \mathbf{F}_{\text{in}} &+ \text{SpatialAttn}(\mathbf{F}_{\text{in}}, \mathbf{f}_{\text{txt}}) + \lambda_1 \cdot \text{SpatialAttn}(\mathbf{F}_{\text{in}}, \mathbf{f}_{\text{aud}}) \\ &+ \lambda_2 \cdot \text{SpatialAttn}(\mathbf{F}_{\text{in}}, \mathbf{f}_{\text{img}}),\end{aligned}$$

where λ_1, λ_2 are learnable fusion weights.

3.3. Motion Interpolator

After generating T_K keyframes, we synthesize intermediate frames using a motion interpolator. Unlike uniform interpolation [1, 24] that predicts between first/last frames, we adapt our keyframe generator to condition on all generated keyframes via masked frame conditioning. We use FreeNoise [15] to generate all T frames in a single pass. Details are in Appendix C.

4. Experiments

4.1. Implementation Details

Datasets. We evaluate on AVSync15 [26], Landscapes [12], and TheGreatestHits [14]. AVSync15 contains 15 activity classes with synchronized audio-video. Landscapes features natural scenes with ambient sounds. TheGreatestHits contains percussive hitting sounds aligned with motions. We use 2-second clips at 24fps (48 frames) resized to 320×512, with $T_K = 12$ keyframes.

Baselines. We compare with: (1) **T+A**: TPoS [10], TempoToken [25]; (2) **I+T**: DynamiCrafter [24], I2VD [26]; (3) **I+T+A**: CoDi [21], AADiff [11], AVSyncD [26].

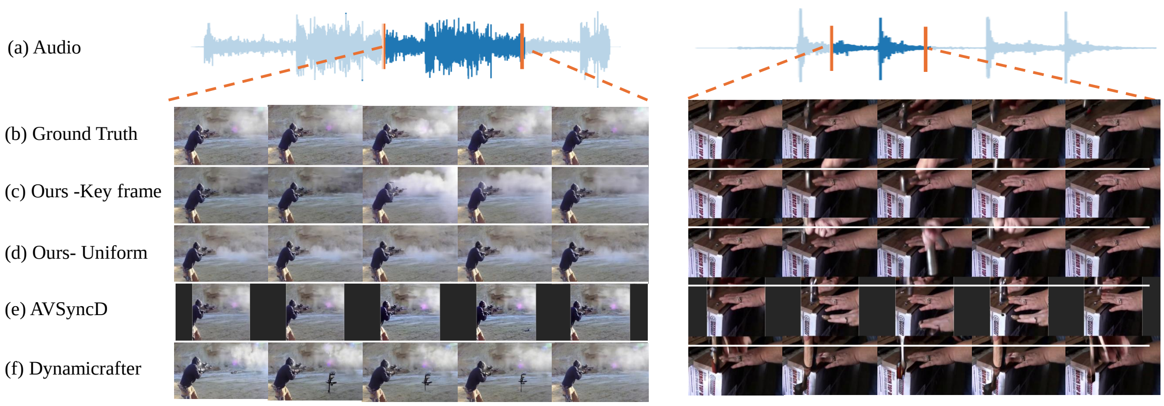
Metrics. We use the Frechet Image Distance (**FID**) [7] and Frechet Video Distance (**FVD**) [23] to evaluate the visual quality and temporal coherence of synthesized videos. We evaluate the audio sycronization with the generated video by **RelSync** and **AlignSync** proposed by Zhang et al. [26].

4.2. Quantitative results

Table 1 presents results on three datasets. On AVSync15, our KeyVID achieves the highest AlignSync (24.09) and RelSync (48.30) scores, demonstrating the effectiveness of our keyframe selection strategy in capturing crucial dynamic moments. Our method also achieves competitive visual quality (FID=11.0, FVD=262.3), outperforming AVSyncD. On Landscapes, which has less dynamics and is used for evaluating visual quality, our method achieves the lowest FVD score (391.09). On TheGreatestHits, featuring distinct percussive audio events, our approach achieves the best performance across all metrics, with notable improvements over AVSyncD.

Input	Model	AVSync15				Landscapes				The Greatest Hit			
		FID↓	FVD↓	AlignSync↑	RelSync↑	FID↓	FVD↓	AlignSync↑	RelSync↑	FID↓	FVD↓	AlignSync↑	RelSync↑
T+A	TPoS [10]	13.5	2671.0	19.52	42.50	16.5	2081.3	23.12	48.15	33.85	3327.90	21.48	44.90
	TempoToken [25]	12.2	4466.4	19.74	44.05	16.4	2480.0	24.21	48.65	25.90	3300.53	21.56	45.38
I+T	I2VD [26]	12.1	398.2	21.80	43.92	16.7	539.5	24.74	49.89	9.1	425.0	22.05	44.58
	DynamiCrafter [24]	11.7	400.7	21.76	43.68	23.51	445.8	24.17	49.63	12.4	337.71	22.82	45.85
I+T+A	CoDi [20]	14.5	1522.6	19.54	41.51	20.5	982.9	22.63	45.48	21.78	1336.00	22.30	45.35
	TPoS [10]	11.9	1227.8	19.67	39.62	16.2	789.6	23.51	47.05	28.43	1370.57	22.04	45.55
	AVSyncD [26]	11.7	349.1	22.62	45.52	**16.2**	415.2	**24.82**	49.93	**8.7**	249.3	22.83	45.95
	KeyVID (Ours)	**11.00**	**262.3**	**24.08**	**48.33**	23.28	**391.0**	24.35	**49.95**	12.1	**202.1**	**22.91**	**46.03**
Static		-	1220.4	21.83	43.66	-	1177.5	25.79	51.59	-	348.9	24.36	48.73
Groundtruth		-	-	25.04	50.00	-	-	25.01	50.00	-	-	25.02	50.00

Table 1. Performance on *AVSync15* and *Landscapes*.



(1) Machine gun shooting (2) Hammering

Figure 4. Qualitative comparison of KeyVID and baseline methods. We crop a key motions on audio waveform in (a) and the corresponding ground truth video in (b) as references and compare the generated video clip between models from (c) to (f). Compared with other models, our KeyVID with keyframe awareness in (c) have better alignment of the motion peaks with audio signals, for example, hitting the hammering, smoke in gun shooting.

4.3. Visualization

Figure 4 presents qualitative comparisons between our method and baseline approaches. Our keyframe-aware approach more accurately captures motion peaks that align with audio events, such as the exact moment of impact in hammering or the smoke in gun shooting. This demonstrates the effectiveness of keyframe-aware training across both high- and low-intensity motion scenarios.

4.4. Ablation Study

We conduct ablation studies to validate the effectiveness of keyframe awareness, as shown in Tab. 2. Specially, we train a variant of the KeyVID model using uniformly sampled 12 frames. Since our method generates high-frame-rate videos (48 frames/2s), we evaluate under two settings: (1) downsample our output to 12 frames to compare with the baseline's 12 uniform frames; (2) interpolate the baseline's 12 frames to 48 using the same method from Sec. 3.3 and evaluate on 48 frames. KeyVID consistently outperforms uniform sampling in both settings, with notable improvements in synchronization metrics (AlignSync and RelSync). These results support our hypothesis that selecting keyframes based on audio and motion cues enhances temporal alignment between audio events and visual dynamics.

Method	FID↓	FVD↓	AlignSync↑	RelSync↑
Evaluate on 12 frames				
KeyVID	**11.00**	**262.34**	**24.08**	**48.33**
Uni. Frame	11.01	273.40	23.53	47.23
Evaluate on 48 frames				
KeyVID	**4.83**	**335.68**	**24.08**	**48.37**
Uni. Frame	4.90	337.10	23.96	48.09

Table 2. Ablation study comparing keyframe-based generation with uniform sampling. KeyVID achieves better performance in both audio synchronization and visual quality with keyframes.

5. Conclusion

We introduce a keyframe-aware, audio-synchronized visual animation model that improves video quality and audio alignment, especially under dynamic motion. Our approach first detects keyframes locations from audio input, then generates them with video diffusion model, and then interpolates intermediate frames for smooth, high-frame-rate output with low memory cost. Experiments on multiple datasets show significant gains in both visual quality and synchronization.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2, 3

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 2

[3] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 1, 2

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2

[6] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 2

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[8] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[9] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[10] Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7822–7832, 2023. 2, 3, 4

[11] Seungwoo Lee, Chaerin Kong, Donghyeon Jeon, and Nojun Kwak. Aadiff: Audio-aligned video synthesis with text-to-image diffusion. *arXiv preprint arXiv:2305.04001*, 2023. 2, 3

[12] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *European Conference on Computer Vision*, pages 34–50. Springer, 2022. 2, 3

[13] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *ArXiv*, 2024. 2

[14] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 3

[15] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 3

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[17] Ludan Ruan, Yunzhi Ma, Hongjie Yang, Haoxian He, Bing Liu, Jianlong Fu, Nenghai Yuan, Qin Jin, and Bing Guo. 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 14012–14021, 2023. 2

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

[19] Xusen Sun, Longhao Zhang, Hao Zhu, Peng Zhang, Bang Zhang, Xinya Ji, Kangneng Zhou, Daiheng Gao, Liefeng Bo, and Xun Cao. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. *arXiv preprint arXiv:2312.01841*, 2023. 2

[20] Zhaohan Tang, Zhilin Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 4

[21] Zhaohan Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. CoDi: Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023. 2, 3

[22] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[23] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 3

[24] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 1, 2, 3, 4

[25] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6639–6647, 2024. 3, 4

[26] Lin Zhang, Shentong Mo, Yijing Zhang, and Pedro Morgado. Audio-synchronized visual animation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1, 2, 3, 4