Animate and Sound an Image

Supplementary Material

A. Additional Implementation Details

Data Processing. In our implementation, video is set to a duration of 2 seconds with a frame rate of 7, while audio preprocessing is consistent with AudioLDM2 [2]. Given that video lengths in all three datasets exceed 2 seconds, we randomly sample a 2-second clip from each video during the training phase. During the inference phase, we randomly sample a 2-second clip from each video in the AVSync15 and Landscape test sets, using the first frame as input for our model and all baselines. For the Greatest Hits dataset, due to many videos exceeding 10 seconds, we segment each video into 10-second intervals, sample a 2-second clip from each segment, and use the first frame as the final test set.

Expert models, model size, frozen layers: Components are tabulated as follows. Joint Block without Expert layer and the A-Expert remain fully trainable due to a small parameter size, ensuring training efficiency. The V-Expert, with a large parameter size, is only trainable on its temporal layers (0.4B), a strategy proven effective in VideoLDM. The video/audio VAE, vocoder and CLIP are frozen in line with standard LDM.

Modules	Param.	Frozen
VideoVAE	97.7M	✓
AudioVAE	55.4M	✓
AudioVocoder	55.3M	✓
CLIP+proj	632M	✓
A-Expert	185M	train
V-Expert	1.5B	partial
JointBlock w/o Expert	77.7M	train

Training and Inference. All the training is done within 4 H100 gpus. We conducted training and testing for each dataset separately due to domain specificity. To account for generation randomness, both the baselines and our JointDiT generate 5, 3, and 3 sounding videos respectively for each test image in the AVSync15, Greatest Hits, and Landscape datasets for evaluation.

Layer Matching: JointDiT accommodates expert models with varying layer configurations by mapping an expert's layer index l_e to JointDiT's block index l_j as follows: $l_j = \begin{cases} \lceil l_e/B \rceil, & \text{if } l_e \leq \lfloor N/2 \rfloor \\ \lceil \lfloor N/2 \rfloor / B \rceil + 1, & \text{otherwise} \end{cases}$ where N the layers of DiT model (blocks of UNet model), and B ($1 \leq B < \min(\lfloor N_{\text{video}}/2 \rfloor, \lfloor N_{\text{audio}}/2 \rfloor)$) is a hyperparameter; smaller B increases the number of joint blocks. In our setting with $N_{\text{video}} = 8$, $N_{\text{audio}} = 9$, and $B = 1$, this results in one input block, three joint blocks, and one output block.

Conditioning Input Methods: For V-Expert, we follow SVD by 1) encoding the image with VAE, duplicating and concatenating it with video latents (noise) along channel dimension; 2) extracting CLIP embeds for the cross-attention. For A-Expert, we modify AudioLDM2's by encoding the image with CLIP, adding it with the timestep embeds.

B. Evaluation Metric Details

B.1. Audio-Video Synchronization Metric

The Audio-Video Synchronization Metric aims to provide a score representing the synchronization degree between a given pair of audio and video. However, the accurate assessment of this synchronization remains a challenge. Existing methods fall into two categories: 1) one relies on a binary classifier trained on curated paired data, consisting of synchronized and unsynchronized examples [3, 6], and 2) the other is based on the matching of audio and video peaks, exemplified by the widely-used AV-Align [5]. The former is constrained by the quality of curated paired data and the strategy for curating positive (synchronized) and negative (unsynchronized) examples, while the latter is limited by the imprecision of calculating the Intersection over Union (IoU) on peaks [5].

We propose an enhancement to AV-Align. Inspired by Wang et al. [4], we treat the video flow features and the audio onset detection curve as two sequences of different lengths, which represent video and audio dynamics respectively. We employ Dynamic Time Warping (DTW), a method designed to measure the match degree between two sequences of different lengths, to calculate the DTW distance between the video flow sequence and the audio onset detection sequence. To mitigate the impact of signal spikes and absolute numerical differences between modalities, we normalize each signal and set parts below 0.2 to 0 to filter out noisy spikes. The resulting DTW distance forms the improved AV-Align metric, noted as AV-Align* in this section, further optimizing the measurement of audio-video synchronization.

We evaluated the above synchronization metrics on the same five settings (SVD+AudioLDM-v, SVD+SeeingHearing, AudioLDM-v +AVSyncD, CoDi, and JointDiT). The correlation between each group of evaluation results and human evaluation scores on five settings was calculated to assess their alignment with human perception of synchronization. As Table 1 shown, the original AV-Align and classifier methods yield lower correlation coefficients 0.164 and 0.206, whereas the im-

	A	B	C	D	E	Corr. ↑
Human Rating ↑	1.25	1.21	1.3	0.87	1.51	1.00
AV-Align ↓	0.357	0.502	0.573	0.435	0.36	0.164
classifier ↑	0.488	0.5	0.511	0.507	0.514	0.206
AV-Align* ↓	1.352	1.569	1.285	1.772	1.296	**0.896**

Table 1. Comparison of the alignment between different synchronization metrics and human perception of synchronization. The columns A-E correspond to five settings in our main text: A: SVD+AudioLDM-v, B: SVD+SeeingHearing, C: AudioLDM-v+AVSyncD, D: CoDi, and E: JointDiT. The column 'Corr.' represents the absolute value of the correlation coefficient between the synchronization scores from each metric and the human-rated synchronization scores for settings A-E. 'Corr.' serves as a measure of the alignment of each metric with human perception of synchronization. Higher Corr. indicates better alignment with human perception of synchronization for a metric.

proved AV-Align* achieves a higher correlation coefficient of 0.896, aligning better with human perception. It's note that we followed Zhang et al. [6] by using the same classifier trained on 2-second video-audio pairs and applied the power exponent normalization method in their work.

Finally, we adopt the enhanced metric, AV-Align*, as our AV-align metric showed in the main text for determining all synchronization scores within our study. It's noteworthy that, although this metric has significantly improved alignment with human perception of synchronization compared to other metrics, it is not yet a perfect indicator. The task of achieving perfect alignment with human perception remains an area for future work.

B.2. Video Dynamic Metrics

Motion Score. Given that image-to-video models often generate static videos [7], we also evaluated the dynamism of the videos in our I2SV task using a motion score [7]. Following the same calculation in SVD [1], the motion score is the sum of the average optical flow values between each pair of adjacent frames, serving as the motion score for the video. This metric is used to quantify the extent of video dynamics introduced by different methods.

C. Additional Case Studies

Figure 1 provides a more extensive comparison between vanilla-CFG and JointCFG*. The top two examples illustrate that our JointCFG* can generate more dynamic striking motions, a case that proves challenging for vanilla-CFG. The bottom two examples show that JointCFG* can maintain the quality of dynamic scenes, such as a rooster raising its head, with clearer details in the rooster's head compared to the vanilla-CFG approach. We posit that JointCFG* can further circumvent generated samples where the original model struggles, such as generating high-dynamic visuals or maintaining quality during dynamic visual generation.

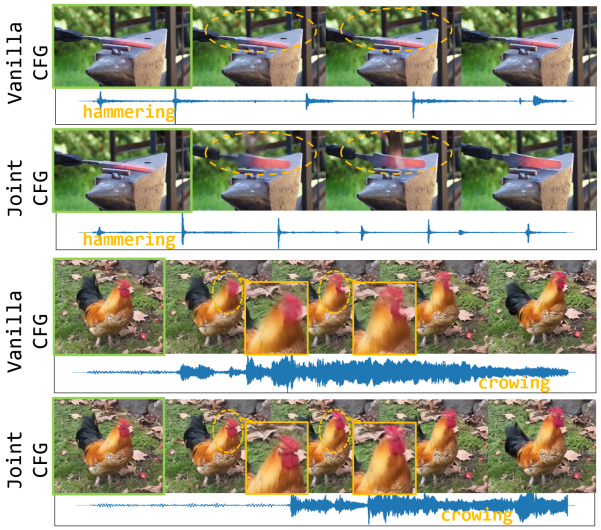

Figure 1. More comparison of different guidance techniques. JointCFG(*) guidance exhibits superior dynamic visual quality.

We believe JointCFG* can exploit the *bad versions* composed of some sub-models to amplify poorly estimated areas in the entire distribution, and then use the comparison in CFG to avoid these poorly estimated areas, achieving boosted performance (better generation quality).

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. 2

[2] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. 1

[3] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2023. 1

[4] Xihua Wang, Yuyue Wang, Yihan Wu, Ruihua Song, Xu Tan, Zehua Chen, Hongteng Xu, and Guodong Sui. Tiva: Time-aligned video-to-audio generation. In *ACM MM*, 2024. 1

[5] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:6639–6647, 2024. 1

[6] Lin Zhang, Shentong Mo, Yijing Zhang, and Pedro Morgado. Audio-synchronized visual animation. In *ECCV*, 2024. 1, 2

[7] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. In *NeurIPS*, 2024. 2