

Supplementary Materials for TiVA

In this supplementary material, we detail previous and our proposed synchronization metrics, alongside an experimental analysis to ascertain their alignment with user-perceived quality.

1 Synchronization Metrics

Synchronization metrics can be categorized by their comparative modalities: 1) visual-to-generated audio, and 2) ground-truth audio-to-generated audio.

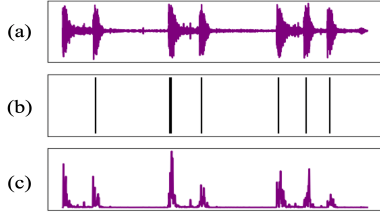


Figure 1: Three audio representations: (a) Waveform; (b) Onset signals; (c) Onset Detection Function (ODF) curve [1].

1.1 Visual-Audio Synchronization Metrics

We investigate two primary metrics for measuring visual-audio synchronization:

- **Align Acc**, introduced by DiffFoley [3], utilizes a trained network to embed both audio and video, and subsequently adjudicates synchronization via a binary score (0/1) output.
- **Temporal Offset (TO)**, from CondFoleyGen [2], categorizes the offsets between audio-visual pairs into discrete values ranging from -2 to 2, with an increment of 0.2. It also relies on an additional trained classifier for embedding and outputting a score.

Both metrics depend on an additional embedding model, which needs extra training costs and requires annotated data, limiting their performance due to the scarcity of labeled data.

1.2 Audio-Audio Synchronization Metrics

According to the representations of audio, the audio-audio synchronization metrics can be further categorized into two classes: *Onset-based Metrics*. These metrics facilitate direct comparisons between the generated and ground truth audio by extracting onsets (Figure 1 (b)), obviating the need for additional network training:

- **Onset Acc** and **Onset Sync AP**, from CondFoleyGen [2], quantify the detection accuracy of onsets and the average precision of synchronization with the ground truth onsets, respectively.

Although these metrics automate calculation, their granularity is limited by binary onsets, resulting in a loss of temporal details.

Onset Detection Function Curve-based Metrics. For a more fine-grained evaluation of synchronization beyond binary onsets, these metrics compare the Onset Detection Function (ODF) curves of the generated and ground-truth audio, as ODF curves capture temporal variations in sound energy, as shown in Figure 1 (c):

- **DTW-dis** treats ODF curves as time series, uses dynamic programming to establish an optimal alignment, and computes the post-alignment Euclidean distance.
- **W-dis** considers ODF curves as unidimensional distributions, quantifying the minimal cost to morph one distribution into another while accounting for distributional distance and mass.

These metrics, requiring no additional network training, more accurately reflect temporal variations in alignment with human perception, as evidenced by our subsequent experiments.

2 Consistency Analysis to Human Evaluation

We perform both human and automatic assessments to benchmark synchronization metrics across various V2A models.

We curate a dataset of 100 samples across 20 categories from the VGGSound test set, selecting videos with varied content and distinct audio-visual sync for metric calibration. We evaluate four models: SpecVQGAN, DiffSound-V (the video-adapted version of DiffSound [4]), DiffFoley, and TiVA with t_a (our method’s upper bound). In the human evaluation, participants were presented with the original and four anonymized reassembled videos, each with audio from a different model, in a random sequence. Ten participants rated each sample on overall quality (Overall) and synchronization (SyncScore) using a 0-5 Likert scale. We compute the mean of these scores and their 95% confidence intervals, as displayed in Table 1. Then, we assess the four models using six automatic temporal synchronization metrics: *Onset Acc*, *Onset Sync AP*, *Align Acc*, *Temporal Offset (TO)*, *DTW-dis*, and *W-dis*, with results in the same Table 1.

Table 1: Comparative evaluation of four models using automatic synchronization metrics and human judgments on curated 100-sample dataset.

Method	Overall \uparrow	SyncScore \uparrow	Onset Acc \uparrow	Onset Sync AP \uparrow	Align Acc \uparrow	TO \downarrow	DTW-dis \downarrow	W-dis \downarrow
SpecVQGAN	1.43 \pm 0.12	1.36 \pm 0.13	0.31	0.64	0.44	1.24	2.93	5.48
DiffSound-V	2.32 \pm 0.10	2.28 \pm 0.13	0.32	0.60	0.81	1.20	2.53	3.64
DiffFoley	2.62 \pm 0.11	2.66 \pm 0.14	0.57	0.30	0.83	1.05	2.37	3.47
TiVA (w/ t_a)	3.20 \pm 0.10	3.41 \pm 0.15	0.57	0.65	0.71	0.77	2.31	2.28

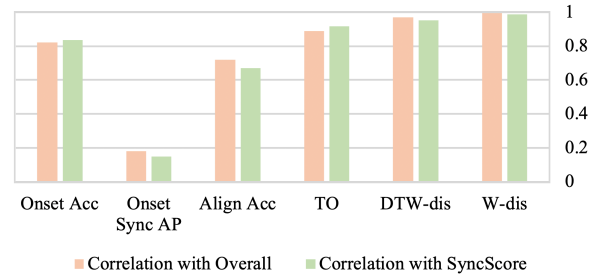


Figure 2: The absolute value of the correlation between automatic and manual evaluation results.

The absolute values of Pearson correlation between automatic and human evaluation are shown in Figure 2. The results indicate that *Onset Sync AP* poorly correlates with human perception of synchronization, and *Onset Acc* and *Align Acc* fail to effectively differentiate models with significant differences in human evaluations. The ODF-curve-based metrics, *W-dis* and *DTW-dis*, exhibit the strongest correlation with human perception, with *W-dis* achieving a 99% correlation with both scores. Consequently, we adopt *W-dis* and *DTW-dis* as our principal synchronization metrics. We also commit to sharing the evaluation codes and dataset with the research community to encourage future exploration.

References

- [1] Juan Pablo Bello, Laurent Daudet, Samer A. Abdallah, Chris Duxbury, Mike E. Davies, and Mark B. Sandler. 2005. A Tutorial on Onset Detection in Music Signals. *IEEE Trans. Speech Audio Process.* 13, 5-2 (2005), 1035–1047. <https://doi.org/10.1109/TSA.2005.851998>
- [2] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. 2023. Conditional Generation of Audio from Video via Foley Analogies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2426–2436. <https://doi.org/10.1109/CVPR52729.2023.00240>
- [3] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. 2023. Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models. *CoRR* abs/2306.17203 (2023). <https://doi.org/10.48550/ARXIV.2306.17203> arXiv:2306.17203
- [4] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2023. DiffSound: Discrete Diffusion Model for Text-to-Sound Generation. *IEEE ACM Trans. Audio Speech Lang. Process.* 31 (2023), 1720–1733. <https://doi.org/10.1109/TASLP.2023.3268730>