Co-Speech Gesture Video Generation with Implicit Motion-Audio Entanglement

Supplementary Material

6. Data Processing

Video Filtering. Some of the collected video candidates may not meet the high-quality standards required for co-speech gesture video generation. For instance, certain videos may feature multiple individuals or exhibit significant scene changes across frames. To handle this, we first segment the videos into clips using SceneDetect[2], ensuring that videos with different scenes are separated. Next, we filter out multi-person videos by employing TalkNet [34], a speaker diarization model that detects and distinguishes different speakers in a video. By using TalkNet, we ensure that the remaining videos contain only single-person scenes. Finally, we use MediaPipe [25] to detect human faces and discard videos with low face detection confidence, which are typically videos featuring side views of faces. Additionally, we retain only video clips longer than 3 seconds, as shorter clips are unlikely to contain meaningful gestures.

Data Annotation. After obtaining the video clips from the previous stage, we annotate our data as follows. First, we extract the audio from the videos using ffmpeg. Next, SHOW [48] method is applied to reconstruct the holistic whole-body mesh, i.e., the SMPL-X motion parameters (including pose and expression). Since the videos are typically rectangular, we need to crop them into square frames. The key challenge during cropping is determining the optimal cropping position. To address this, we render the mesh parameters into mesh frames. Then, we binarize these mesh images to obtain segmentation masks. Using these masks, we compute the largest bounding box of the person in the video, crop the width based on this value, and pad the height accordingly to achieve a square aspect ratio. By ensuring that all frames in the same video use the same bounding box, we maintain a consistent camera position, which is crucial for our task, as modeling camera movement is challenging.

After processing the data, we obtain our final dataset and split it into training and test sets. The detailed statistics are provided in Table. 1.

7. Implementation Details

The training process is divided into two stages. In the first stage, we train only the audio-to-motion branch using audio and motion parameters. This stage runs for 3,000 epochs with a learning rate of 1e-4. The first stage is trained on 8 RTX 8000 GPUs for 1 day with a batch size of 256. In the second stage, we train the entire network for 200 epochs with a learning rate of 1e-5. This stage is trained on 4 A100 GPUs for 4 days with a batch size of 1. The input image resolution is 512x512. The classifier-free guidance (CFG) scale is set to 3.5 for the video diffusion branch.

To improve the model's learning by exposing it to more frames, we replace the 2D-VAE in I2VGen-XL [50] with 3D CV-VAE [51]. In this setup, the frame sequence length F is set to 80 for the audio-to-motion branch. For the video diffusion branch, the input is processed as 64 frames (f), which are reduced to 16 latent frames after passing through the 3D CV-VAE. In addition, the training videos have an FPS of 30, allowing the video diffusion branch to process approximately 2 seconds of video.

For experimental efficiency, we train a separate model for each identity. However, our method fully supports multi-identity training. The reference encoder controls identity appearance, and a one-hot ID embedding can be easily integrated to audio-to-motion branch.

Efficiency Comparison: we compare inference time and memory here. For 1s video (30fps, 512×512) generation on an A100 GPU, S2G takes 4.6s and 3GB, MYA takes 40s and 46GB, while ours takes 13.7s and 21GB. Note that MYA and ours are tested with memory-efficient attention disabled due to environmental issues. Compared to diffusion-based MYA, ours is more efficient due to the lightweight reference encoder and audio-to-motion branch.

8. Metric Implementation

To evaluate the generated videos, we first extract skeleton keypoints using the DWPose framework [45], preserving 12 upper-body keypoints and 21 keypoints for each hand, totaling 54 keypoints. FGD and Diversity are computed using an autoencoder trained on skeleton keypoints from our training dataset, as outlined in [29]. Details regarding autoencoder training and FGD computation are available in their github repository[3]. For Diversity, we adopt the methodology from [52], with implementation details found in their github repository[4].

Additionally, BAS is calculated directly following [20][5]. For FVD, we leverage the I3D classifier [39], which is pretrained on the Kinetics-400 dataset [18]. Further instructions for this metric are available in the associated github repository[6].

[2] https://github.com/Breakthrough/PySceneDetect

[3] https://github.com/ShenhanQian/SpeechDrivesTemplates

[4] https://github.com/Advocate99/DiffGesture/tree/main

[5] https://github.com/google-research/mint

[6] https://github.com/JunyaoHu/common_metrics_on_video_quality

9. More Comparisons with MYA

In the main text, we utilize DiffSHEG [6] to generate motion parameters, render them into pose images, and drive MYA to produce videos. To eliminate the influence of DiffSHEG, we directly use ground truth motion parameters, render them into pose images, and drive MYA for video generation. As shown in Table 4, the GT-driven MYA unsurprisingly achieves the best performance on FGD and BAS. However, our method outperforms it on Diversity and FVD metrics, indicating that our approach generates more diverse gestures than even the ground truth. Moreover, our method produces videos with stable backgrounds and clear hand and finger details, whereas MYA generates videos with unstable backgrounds and distorted hands and fingers, as illustrated in Fig. 7. Furthermore, MYA often memorizes appearance features during training, leading to generated videos that replicate the memorized appearance rather than relying on the reference image. This behavior results in noticeable inconsistencies, further highlighting the limitations of MYA in generating diverse and accurate outputs.

Additionally, to ensure that our performance is not solely attributed to I2VGen-XL [50], we replace the image diffusion model with I2VGen-XL and adjust the dimensions accordingly for video input. As shown in Table 4, this modification yields slightly improved results compared to SD-MYA, though it remains inferior to our approach. Moreover, video diffusion models primarily enhance temporal smoothness rather than addressing hand artifacts, and the pretrained model mainly serves as a better initialization since we employ full-parameter fine-tuning.

10. User Study

To further evaluate the visual performance of our method, we conduct a user study comparing the gesture videos generated by each method and each ablation study. We sample 30 generated videos from our test set for each method, and 20 participants are invited to rank the videos. Participants are asked to evaluate the videos based on four criteria:

Identity Preservation: Evaluates how well the essential characteristics and attributes of the human are maintained across the video.

Visual Quality: Assesses the video's clarity, with higher rankings indicating fewer issues such as blur, noise, and visual degradation.

Temporal Consistency: Measures frame-wise coherence, ensuring the logical progression of motion and visual elements across consecutive frames.

Sound-Video Synchronization: Judges the alignment between speech and gestures, assessing the accuracy of the generated motions.

Participants rank the videos, with rank 1 being the best. In comparison with previous works, the rankings are converted into points: rank 1 is assigned 3 points, rank 3 is given 1 point, and so on. For ablation studies, rank 1 is assigned 5 points, rank 5 is given 1 point, and so on. A higher overall score indicates better performance.

The user study results are presented in Table 5 and Table 6. As shown in Table 5, our method significantly outperforms others across all dimensions, demonstrating its ability to generate gesture videos with superior motion quality and overall visual fidelity. Although MYA achieves a slightly better BAS, it does not affect human perception of synchronization. Table 6 further highlights that our full model achieves state-of-the-art results in all metrics. The model without motion information performs the worst, which is consistent with the objective results shown in Table 3. The model without a reference encoder and the model without first-stage training yield comparable results, indicating that skipping the first-stage training shifts focus to the audio-to-motion branch while reducing the emphasis on the video diffusion branch, thereby degrading the visual quality. The model without slow-fast training achieves the second-best results but still falls short of our full model, demonstrating the effectiveness of our slow-fast training strategy.

We also show the user study interface in Fig. 10.

10.1. Statistical Analysis

Given the limited number of participants, slight differences in rankings may not reliably indicate a significant preference for one method over another. To address this issue, we apply three statistical tests to validate the effectiveness of our user study.

Kruskal–Wallis Test. We use the Kruskal-Wallis test to assess overall differences across multiple groups. This test is particularly robust when dealing with ordinal data and does not require a normal distribution, making it well-suited for our dataset. Since our user study data is ordinal and non-normally distributed, the Kruskal-Wallis test provides a reliable way to evaluate statistical significance. For more details on the calculation, readers can refer to the Wikipedia page[7]. The test outputs a p-value, where a lower value indicates a higher degree of confidence in the observed differences across groups, signifying a stronger statistical significance.

Dunn's Test. Dunn's Test [9] is a post-hoc test used for pairwise comparisons between groups. If the Kruskal-Wallis test indicates a statistically significant difference, Dunn's Test helps identify which specific groups differ from each other.

As shown in Table 7 and Table 8, the p-values are very low and approach zero, indicating substantial overall differences across the groups. To further analyze these differences, we refer to Fig. 8 and Fig. 9 for the results of Dunn's Test. In Fig. 8, all three groups show significant differences,

[7] https://en.wikipedia.org/wiki/Kruskal–Wallis_test

Figure 7. Qualitative comparison with GT-driven MYA. The leftmost image is used as the reference image. Red circles highlight the obvious flaws in MYA. As shown, it struggles with issues such as unstable background, blurry hands and distorted fingers.

Model	FGD ↓	Div. ↑	BAS ↑	FVD ↓
S2G [12]	3.69	180.59	0.7280	816.03
MYA [17] + DiffSHEG [6]	24.24	224.14	0.7452	1823.97
I2VGen-XL [50]+MYA [17] + DiffSHEG [6]	22.68	233.72	0.7427	1664.70
MYA [17] + GT	**0.18**	180.92	**0.7542**	841.31
Ours	1.87	**273.72**	0.7445	**681.33**

Table 4. Quantitative comparison with previous works on four objective metrics. Bold text indicates the best performance.

validating the effectiveness of our user study. For instance, our method outperforms S2G by 0.5-0.6 points across all metrics (Table 5), and Dunn's Test confirms that the differences between the two groups are statistically significant, demonstrating that our method is superior and unaffected by the limited sample size. An interesting observation can be seen in Fig. 9, where the model without first-stage training and the model without the reference encoder show no significant difference, as indicated by a p-value of 1. This finding is consistent with the results in Table 6, where columns 1 and 2 yield similar scores. This suggests that it is difficult to determine which model performs better given the limited number of participants. However, this does not undermine the validity of the user study, as our full model demonstrates a statistically significant difference compared to all other incomplete models.

ABX Test. First, we present the statistical analysis comparing our method (Ours) with S2G and MYA, in Table 9. The score summary for these videos is presented in Table 10. Based on the statistical analysis:

1. Significant differences exist across all metrics (all p-values <0.05).
2. Ours performs best: Highest mean scores (2.65-2.75), superior across all metrics vs S2G and MYA, and most consistent performance (lower standard deviation relative to mean).
3. MYA performs worst: Lowest scores (1.10-1.22), largest negative difference vs reference, particularly poor in Identity Preservation (1.10).
4. Metric-specific findings: Identity Preservation shows the largest variance between methods, Sound-Video Sync differences are significant but smaller, and Visual Quality and Temporal Consistency show moderate differences.

To summarize, the ABX test result shows significant differences across all video quality metrics (p-values <0.05), with ours demonstrating superior performance (mean scores 2.65-2.75) and significant improvements over alternatives (t-statistics 15.87-60.97). MYA consistently underperforms (mean scores 1.10-1.22), while S2G falls between these ex-

Model	Preservation ↑	Quality ↑	Consistency ↑	Synchronization ↑
S2G	2.15	2.12	2.18	2.18
MYA	1.10	1.22	1.13	1.13
Ours	**2.75**	**2.66**	**2.69**	**2.70**

Table 5. Quantitative comparison with previous works on four subjective metrics. Bold text indicates the best performance.

Model	Preservation ↑	Quality ↑	Consistency ↑	Synchronization ↑
w/o Ref	2.52	2.49	2.48	2.54
w/o Motion	1.86	1.83	1.89	1.84
w/o First Stage	2.42	2.52	2.49	2.52
w/o Slow-Fast	3.51	3.55	3.48	3.55
Ours	**4.69**	**4.61**	**4.66**	**4.53**

Table 6. Quantitative ablation study on four subjective metrics. Bold text indicates the best performance.

tremes but remains significantly below ours (mean differences 0.50-0.60). These results align with our previous Kruskal-Wallis test, indicating statistically significant differences across all videos and supporting ours as the optimal approach.

Next, we present the ablation study comparing our full method (Ours) with ablated versions: w/o Ref, w/o Motion, w/o First Stage, and w/o Slow-Fast, in Table 11. The score summary for these videos is shown in Table 12. Based on the statistical analysis:

1. Significant differences exist across all metrics (all p-values <0.05).
2. Ours performs best: Highest mean scores (4.54-4.69), superior across all metrics vs w/o Ref, w/o Motion, w/o First Stage, and w/o Slow-Fast, and most consistent performance (lower standard deviation relative to mean).
3. w/o Motion performs worst: Lowest scores (1.83-1.89), largest negative difference vs reference, particularly poor in Identity Preservation (1.86).
4. Metric-specific findings: Identity Preservation shows the largest variance between methods, Sound-Video Synchronization differences are significant but smaller, and Visual Quality and Temporal Consistency show moderate differences.

To summarize, the ABX test result for the ablation study shows significant differences across all video quality metrics (p-values <0.05), with ours demonstrating superior performance (mean scores 4.54-4.69) and significant improvements over ablated versions (t-statistics 17.63-66.32). w/o Motion consistently underperforms (mean scores 1.83-1.89), while w/o Slow-Fast performs closest to ours but remains significantly below (mean differences 0.99-1.19). These results highlight the importance of each component in our method, particularly motion modeling, and support ours as the optimal configuration.

11. Discussion on Lip Synchronization

Our method demonstrates strong performance in mitigating hand artifacts and aligning gestures with audio, yet it struggles to achieve precise lip synchronization in co-speech gesture video generation. Effective lip sync typically demands either large-scale datasets—such as the 2.2K hours used in VLOGGER [7]—or specialized designs, like the latent facial expression space in EmoPortraits [8]. Other approaches also achieve lip synchronization either by training on large-scale datasets using diffusion models [35, 42, 43] or by incorporating specialized lip synchronization modules [28, 47]. However, our dataset is too limited in size to support such training, and the lip region constitutes only a small fraction of the RGB feature space. This makes it difficult to attain accurate lip sync using diffusion learning alone, as the subtle details required for lip movements are not sufficiently captured.

Although our audio-to-motion branch incorporates expression parameters, these form just a minor component of the overall motion parameters and are not specifically engineered for lip synchronization in the RGB space. These parameters are fed into the video diffusion branch via cross-attention, but without dedicated modules or loss functions targeting lip movements, control over lip sync remains implicit and weak rather than explicit and robust. This lack of tailored design exacerbates the challenge of achieving precise lip alignment.

11.1. Audio-to-Motion Generation Quality

To address potential concerns that the lip synchronization issues arise from inadequate audio-to-motion parameters, we train our model on the TalkShow dataset [48] for a fair comparison with TalkShow method. Our method outperformed TalkShow method across both face and body metrics, as shown in Table 13.

These results indicate that our audio-to-motion parame-

	Preservation	Quality	Consistency	Synchronization
P-Value	9.37×10^{-256}	1.73×10^{-193}	1.43×10^{-233}	6.65×10^{-235}

Table 7. Kruskal-Wallis Test results on the user study of comparisons with previous works.

	Preservation	Quality	Consistency	Synchronization
P-Value	7.37×10^{-310}	2.33×10^{-297}	5.37×10^{-296}	1.12×10^{-275}

Table 8. Kruskal-Wallis Test results on the user study of ablation analysis.

ters are sufficiently accurate and not the root cause of the lip synchronization shortcomings. In addition, unlike TalkShow, which renders mesh outputs, our model generates RGB features, introducing additional complexity due to the need to handle detailed textures and appearances. This distinction suggests that the lip synchronization challenge is more closely tied to the limited dataset size and the absence of specific lip synchronization mechanisms rather than deficiencies in the audio-to-motion pipeline.

12. Future Work

Although our method demonstrates strong performance in co-speech gesture video generation, there is significant potential for further improvement. Below, we outline several key areas for future exploration.

Larger Dataset. Although we introduce a new large-scale dataset, it only includes four identities and 33 hours of video. A more comprehensive benchmark is needed for this task. A key question is determining the dataset size required for each identity to accurately generate high-quality videos that replicate individual gesture styles.

Advanced Attention Mechanism. Our current approach uses a basic cross-attention mechanism to connect the audio-to-motion and video diffusion branches. Future work could explore more advanced attention mechanisms to better capture and represent expression and pose, thereby enhancing the motion information for the video diffusion process.

Imbalanced Identities. Despite containing only four identities, our dataset suffers from an imbalance in data distribution across subjects. This issue requires deeper analysis and effective solutions to ensure a balanced representation of model training.

Diverse Video Generation. While our current method is limited to single-person videos, fixed backgrounds, and front-view perspectives, future work will focus on generating diverse videos that feature multi-person interactions, dynamic viewpoint changes, and adaptable backgrounds. This would enable the creation of more versatile and realistic scenarios, greatly enhancing the applicability and robustness of the approach.

13. Ethical Impacts

Co-speech gesture video generation, which synthesizes human-like gestures aligned with speech, raises several ethical concerns. One major risk is its potential misuse in deepfake technology, which can spread misinformation and deceive audiences. Additionally, replicating culturally specific gestures without proper context may lead to miscommunication or cultural insensitivity. Safeguarding individual privacy and ensuring informed consent is essential, especially when using real individuals' data. Furthermore, over-reliance on this technology in human-computer interactions may diminish the richness of authentic non-verbal communication. To mitigate these risks, ethical guidelines and robust safeguards must be established to promote responsible development and use.

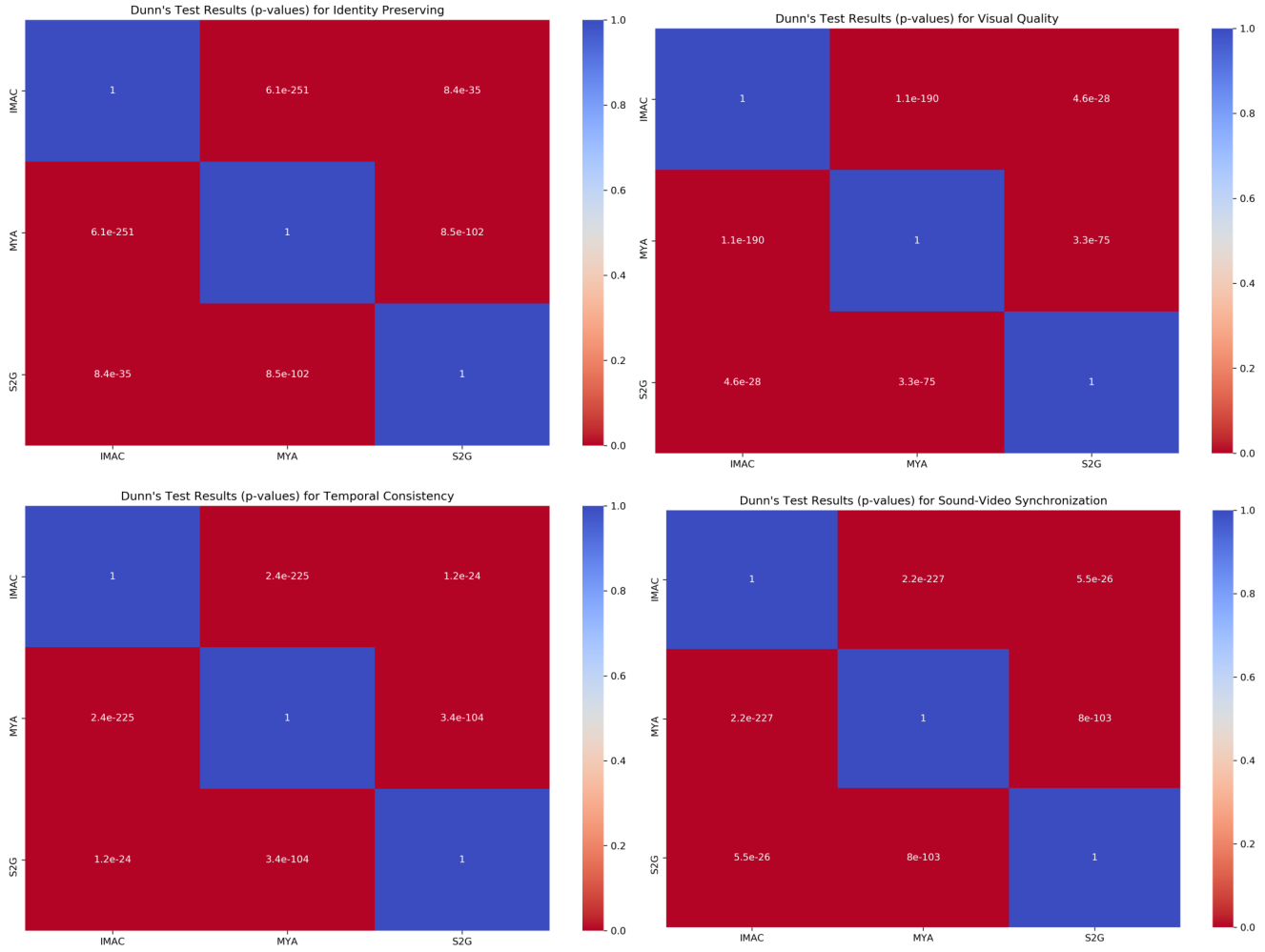

Figure 8. Heat map of pairwise comparisons using Dunn's test for the user study comparing with previous works. Warmer colors (closer to red) indicate greater statistical significance in the differences between models, while cooler colors (closer to blue) denote lower statistical significance. This visualization facilitates the quick identification of the most and least distinct model pairs.

Metric	Compared Video	Mean Difference	T-Statistic	P-Value	Significant
Identity Preserving	S2G	0.603	20.35	1.14×10^{-78}	True
	MYA	1.647	60.97	0.00	True
Visual Quality	S2G	0.539	15.88	2.25×10^{-51}	True
	MYA	1.434	41.10	2.79×10^{-226}	True
Temporal Consistency	S2G	0.504	15.82	4.99×10^{-51}	True
	MYA	1.560	52.00	1.35×10^{-301}	True
Sound-Video Synchronization	S2G	0.518	16.11	1.01×10^{-52}	True
	MYA	1.567	54.12	1.52×10^{-315}	True

Table 9. Statistical Analysis for Video Comparisons (Reference: Ours)

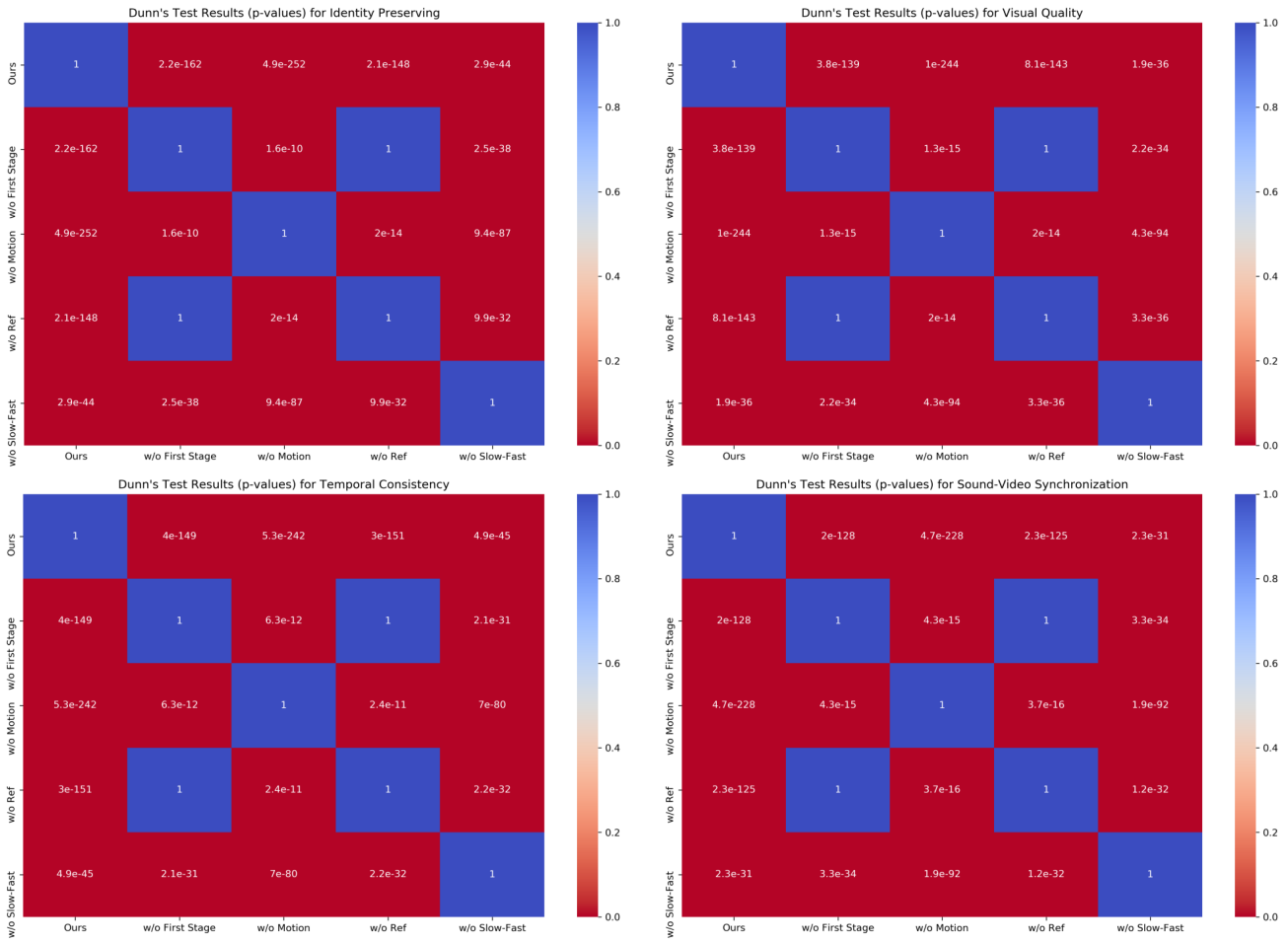

Figure 9. Heat map of pairwise comparisons using Dunn's test for the user study in the ablation study. Warmer colors (closer to red) indicate greater statistical significance in differences between models, while cooler colors (closer to blue) represent lower statistical significance. This visualization allows for quick identification of the most and least distinct model pairs.

Video	Identity Preserving	Visual Quality	Temporal Consistency	Sound-Video Synchronization
S2G	2.147 ± 0.447	2.119 ± 0.514	2.184 ± 0.474	2.177 ± 0.502
MYA	1.103 ± 0.342	1.223 ± 0.549	1.128 ± 0.401	1.128 ± 0.379
Ours	2.750 ± 0.542	2.658 ± 0.620	2.688 ± 0.588	2.695 ± 0.574

Table 10. Score Summary for Videos (Comparison with S2G and MYA)

Metric	Compared Video	Mean Difference	T-Statistic	P-Value	Significant
Identity Preserving	w/o Slow-Fast	1.175	22.89	7.03×10^{-96}	True
	w/o Ref	2.166	51.65	2.40×10^{-302}	True
	w/o Motion	2.827	66.32	0.00	True
	w/o First Stage	2.266	35.77	3.73×10^{-189}	True
Visual Quality	w/o Slow-Fast	1.064	20.95	7.23×10^{-83}	True
	w/o Ref	2.125	49.22	9.98×10^{-286}	True
	w/o Motion	2.785	65.84	0.00	True
	w/o First Stage	2.097	31.74	1.82×10^{-159}	True
Temporal Consistency	w/o Slow-Fast	1.185	22.19	3.93×10^{-91}	True
	w/o Ref	2.187	50.02	2.93×10^{-291}	True
	w/o Motion	2.770	64.53	0.00	True
	w/o First Stage	2.171	33.91	2.01×10^{-175}	True
Sound-Video Synchronization	w/o Slow-Fast	0.985	17.63	8.07×10^{-62}	True
	w/o Ref	1.990	41.01	1.62×10^{-227}	True
	w/o Motion	2.691	59.07	0.00	True
	w/o First Stage	2.014	29.79	4.05×10^{-145}	True

Table 11. Statistical Analysis for Video Comparisons (Reference: Ours)

Video	Identity Preserving	Visual Quality	Temporal Consistency	Sound-Video Synchronization
w/o Ref	2.521 ± 0.851	2.490 ± 0.883	2.476 ± 0.890	2.545 ± 0.941
w/o Motion	1.860 ± 0.870	1.829 ± 0.859	1.893 ± 0.868	1.844 ± 0.850
w/o First Stage	2.420 ± 1.423	2.517 ± 1.491	2.491 ± 1.434	2.521 ± 1.471
w/o Slow-Fast	3.512 ± 1.109	3.550 ± 1.092	3.478 ± 1.156	3.550 ± 1.152
Ours	4.687 ± 0.540	4.614 ± 0.544	4.663 ± 0.557	4.535 ± 0.688

Table 12. Score Summary for Videos (Ablation Study)

Method	Face Metrics			Body Metrics	
	JawL1	LandmarkL1	LVD	LVD	FGD
TalkShow	0.00158	0.1553	0.0278	0.0229	2.91
Ours	0.00142	0.1485	0.0233	0.0189	2.12

Table 13. Comparison of audio-to-motion metrics between TalkShow and our method on the TalkShow dataset. Note that the metrics differ from those in TalkShow paper, please refer to their github repository (Issue 4) for details.

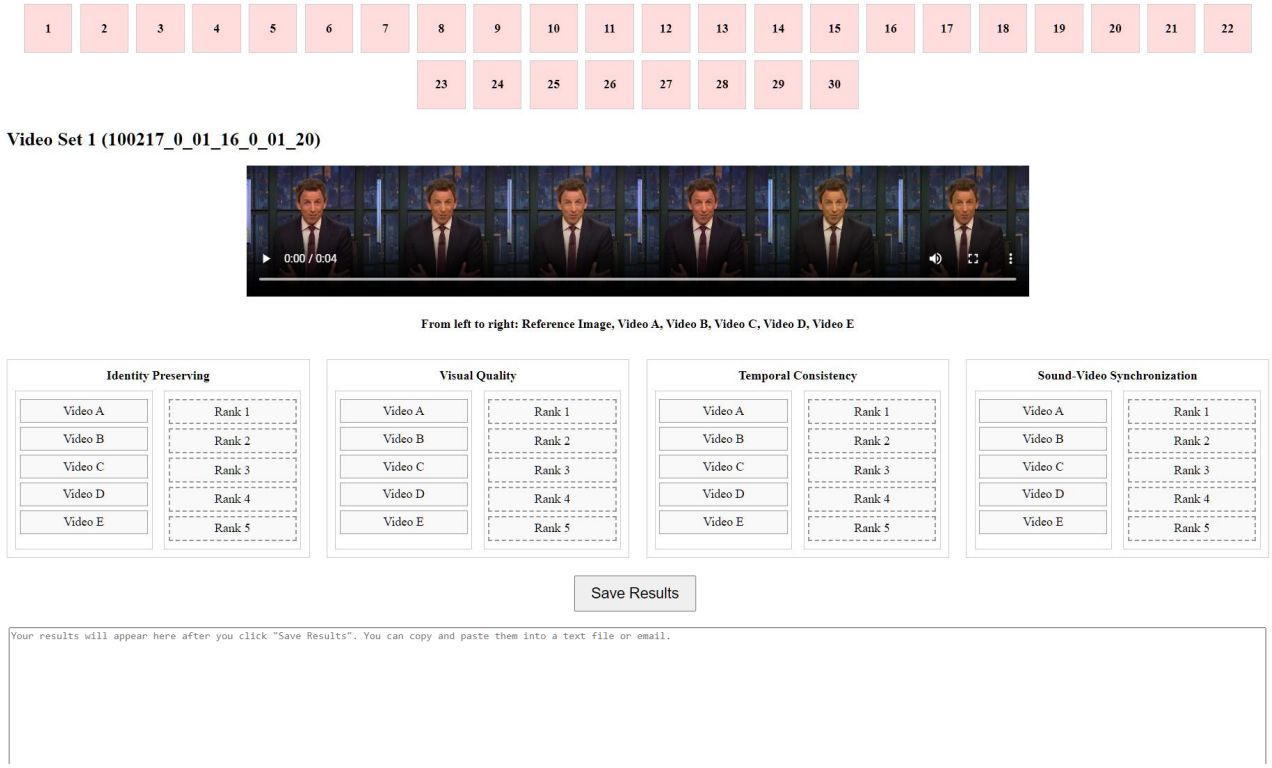

Figure 10. The interface allows users to drag the video ID to the corresponding rank ID, with the option to double-click to cancel the selection. Final rankings are displayed in the result box after clicking ”Save Results.” We also design an interactive window where incomplete tasks are highlighted in red, enabling users to identify and address any unranked videos easily.