# Talking Together: Synthesizing Co-Located 3D Conversations from Audio

Supplementary Material

## 1. Video Results

To demonstrate the effectiveness of our framework, we provide a folder containing 14 distinct conversational sequences sourced from our test set. This can be best viewed by clicking on the `video_results.html` file with your favorite browser (Chrome recommended!) These examples cover diverse dynamics, including rapid turn-taking and overlapping speech. For each sequence, we compare four methods:

- **Ours (Full):** Our complete dual-stream model.
- **Ours (Stage 1):** Ablation without the second-stage high-fidelity lip fine-tuning.
- **DualTalk:** The state-of-the-art dual-speaker baseline.
- **SelfTalk:** Single-speaker baseline applied to each track.

We encourage observation from the following aspects:

- **Listener Responsiveness:** Ours generates natural reactions (nods, eye movements) while listening, whereas SelfTalk often remains frozen.
- **Interaction Coherence:** Ours maintains spatial awareness and mutual gaze, avoiding the disconnected "video conference" feel of DualTalk.
- **Lip-Sync Precision:** Comparing Ours (Full) vs. Ours (Stage 1) highlights the necessity of our second-stage fine-tuning for sharp, accurate articulation.

## 2. Implementation Details

### 2.1. Latent Diffusion Model

Our model is an audio-driven facial animation system based on a latent diffusion model, designed to generate sequences of parameters for a 3D parametric face model, including expression, joint rotations, and root translation. The generation is conditioned on pre-trained audio embeddings and facial identity. The core network is a U-Net with an embedding dimension of 512. This U-Net consists of 2 main blocks, each with 2 subblocks. The second block incorporates self-attention with 8 heads and a temporal stride of 2. The model predicts parameters for a reduced PCA version of the face model, encompassing 63 dimensions for expression, 50 for identity, 12 for joint rotations (representing 4 joints as axis-angle vectors), and 3 for root translation. The model processes sequences of 250 frames. We employ classifier-free guidance with a conditioning signal dropout probability of 0.1 and a guidance scale of 2.0 during inference, which utilizes 8 DDIM sampling steps. The diffusion process is trained to predict the clean data ($x_0$) using a cosine-based logSNR schedule. The model is trained using the Adafactor optimizer with a learning rate of 1e-4, which remains constant after a 2000-step linear warmup, and a batch size of 1024. The loss function is a weighted sum of L2 losses on the face model parameters with specific setup and weights described in the paper.

### 2.2. Few-Shot Spatial Layout Generation

To enable text-driven control over the initial spatial arrangement of the participants, we utilize **Gemini 2.5 Pro** via API. We employ a few-shot prompting strategy to condition the model to output precise 3D translation vectors $t \in \mathbb{R}^3$ (in metric meters) for both Speaker A and Speaker B.

**Prompt Structure.** The prompt consists of three parts: (1) a system instruction defining the coordinate system (where $+Z$ is forward and $+X$ is right) and the origin point (center of the interaction); (2) three in-context examples demonstrating diverse spatial relationships; and (3) the user's input query. We enforce a strict JSON output format to ensuring programmatic parsability.

**Inference Parameters.** We set the temperature to $0.2$ to encourage deterministic and stable coordinate predictions. The maximum output token limit is set to 128.

**Prompt Template.** Below is an example structure of the prompt provided to the LLM. In practice, we manually label 30 examples, and randomly choose 3 as few-shot examples in each inference round.

```
System: You are a 3D scene layout
assistant. Generate 3D head
translation coordinates (in meters)
for two people (A and B) based on
a text description. The center of
the conversation is (0,0,0). Output
ONLY valid JSON.

[Example 1]
Input: "Standing face-to-face in a
normal conversation."
Output: {"A": [0.0, 0.0, -0.5],
"B": [0.0, 0.0, 0.5]}

[Example 2]
Input: "Sitting side-by-side on a
bench watching a game."
Output: {"A": [-0.3, -0.2, 0.0],
"B": [0.3, -0.2, 0.0]}

[Example 3]
Input: "A is whispering into B's
ear."
Output: {"A": [-0.15, 0.0, -0.1],
"B": [0.15, 0.0, 0.1]}
```

```
[User Query]
Input:  "<USER_PROMPT>"
Output:
```

The returned JSON is parsed, and the coordinates are used as the condition $t_A^{(0)}$ and $t_B^{(0)}$ for the diffusion model.

# 3. Evaluation Details

## 3.1. Baseline Details

We provide implementation details for all baselines to ensure fairness and reproducibility. Since these models differ significantly in their input formats and capabilities, we standardize audio processing, temporal alignment, and rendering across all comparisons.

**Audio Processing.** All baselines use the same audio preprocessing pipeline as our method. We apply Looking-to-Listen for audio source separation and WebRTC VAD for frame-level activity detection, producing two separated audio streams. The separated audio is resampled to 16 kHz and converted into Wav2Vec 2.0 features. This ensures identical audio conditions for all models.

**Single-Speaker Baselines.** We include CodeTalker, FaceFormer, SelfTalk, and a strong single-person model trained on our single-speaker dataset. When available, we use official pretrained weights; otherwise, we train using the authors' public code on single-speaker data only. Because these models do not support conversational interaction, we run each participant independently on its separated audio track without any listener modeling. These baselines predict expression and rotation only. For metrics requiring translation, we use the predicted translation from our model to place the generated meshes in a consistent 3D frame. All outputs are resampled or interpolated to 25 fps.

**Dual-Speaker Baseline.** To construct a competitive two-speaker baseline, we use SelfTalk—the strongest single-person performer in our experiments—to produce the primary speaker motion. A listener-generation model is then conditioned on this motion to synthesize the second participant's reactive behavior. This two-stage pipeline approximates interactive motion but lacks true joint reasoning or simultaneous prediction.

**Retrieval-Based Baselines.** We include Listen-R (random retrieval), Listen-A (audio-based retrieval), and Listen-M (motion-based retrieval). For Listen-A, we compute cosine distances between Wav2Vec features of the speaking segment and all segments in the training corpus. For Listen-M, retrieval is based on a joint distance over expression trajectories and head-rotation curves. Retrieved sequences are trimmed or padded to match the speaking segment duration for strict temporal alignment.

**Rendering and Alignment.** All baselines are rendered using the same 3DMM, renderer, and camera configuration as our method to avoid artifacts arising from differences in visualization. For models lacking translation, we apply our predicted translations to maintain a consistent spatial coordinate system. All sequences are normalized to 25 fps before evaluation.

Table 5. Comparison of expression-space mappings. We report cycle-consistency error (lower is better) and perceptual similarity between rendered faces using LPIPS (lower is better) and SSIM (higher is better). Our three-layer MLP significantly outperforms a linear regression baseline.

| **Method** | **Cycle Error** $\downarrow$ | **LPIPS** $\downarrow$ | **SSIM** $\uparrow$ |
|---|---|---|---|
| Linear Map | 0.684 | 0.212 | 0.912 |
| 3-Layer MLP | 0.327 | 0.141 | 0.963 |

## 3.2. Face Model Mapping

To integrate baselines that use a different facial parameterization, we learn a bidirectional translation between the 51-D expression space used by their model and the 63-D expression space used in ours. We construct a paired dataset by sampling around 1,100,000 single-person talking videos in our dataset containing a wide range of facial motions, including challenging and extreme expressions. For each frame, we fit both face models, yielding paired expression vectors. Using this dataset, we train two three-layer MLPs (128 hidden units, ReLU activations) to map between the two expression domains.

**Cycle Consistency.** We assess the quality of the learned mappings through a cycle-consistency evaluation. For an expression vector $x^A$ in the 51-D space, we compute its mapped counterpart $x^B = f(x^A)$, then map it back to $\hat{x}^A = g(x^B)$. The cycle error between $\hat{x}^A$ and the original $x^A$ provides a direct measure of how well the two expression spaces are aligned under the learned transformations. Low cycle error indicates that the mappings preserve the underlying geometry of the expression space and avoid collapsing or distorting important facial degrees of freedom.

**Perceptual Similarity via Rendering.** To validate the perceptual fidelity of the mapping, we render faces driven by the original and cycle-mapped expressions using our 3DMM renderer. We then compute perceptual similarity metrics such as LPIPS and SSIM between the rendered images. These metrics capture subtle visual differences (particularly around the lip, jaw, and eye regions) and provide a direct indication of whether the mapping preserves the visual appearance of expressions when applied to animation.

**Comparison with a Linear Model.** As a baseline, we train a linear regression model on the same paired data. We compare its cycle-consistency error and perceptual similarity metrics with those obtained from our MLP. Results are shown in Tab. 5. Given that cycle error is significantly

smaller than natural expression variance and that the MLP substantially outperforms a linear regressor, we conclude that the mapping is accurate, robust, and conservatively sufficient for fair cross-model evaluation.

### 3.3. Evaluation Metrics

To comprehensively assess the performance of our system in terms of motion realism, interaction coherence, and geometric accuracy, we employ the following set of metrics:

**Fréchet Distance (FD).** FD measures the perceptual realism of the generated outputs by comparing them to ground-truth samples in a deep feature space. In our evaluation, FD is computed on rendered images rather than motion parameters. Each frame is rendered after normalizing translations and rotations to a zero-centered pose. We extract image features using a pre-trained Inception encoder and model their distributions as multivariate Gaussians. FD then quantifies the distance between generated and real distributions, where lower values indicate higher perceptual fidelity.

**Paired Fréchet Distance (P-FD).** P-FD extends FD to assess the quality of dyadic interaction. We render paired frames of the two participants with zero-centered pose normalization and extract their Inception image features. The features of both individuals in each frame are concatenated before computing FD between the generated and ground-truth paired distributions. This evaluates the coherence, synchrony, and interaction consistency of the generated pairwise behaviors.

**Mean Squared Error (MSE) and Vertex MSE (vMSE).** We employ two direct reconstruction metrics.

- **Parameter MSE** computes the average squared difference between predicted animation parameters (expression, rotation, translation) and the ground truth. This measures how well the model reproduces the underlying control signals.
- **Vertex MSE** computes the Euclidean distance between the vertices of the generated 3D mesh and those of the ground-truth mesh. Because it evaluates geometric deformation directly in 3D space, vMSE correlates more strongly with perceptual differences on the facial surface.

**Region- and Role-Specific Evaluation.** To provide a fine-grained understanding of model performance, we compute the above metrics over specific subsets:

- **Speaker vs. Listener:** These two conversational roles exhibit distinct motion characteristics. Speaker metrics emphasize articulation quality and lip-synchronization, while Listener metrics capture non-verbal cues such as nods, blinks, and backchanneling behaviors.
- **Facial Regions (Lip, Eye, Global):** We evaluate distinct facial regions to isolate component behavior. Lip error reflects alignment with speech; Eye error captures the effectiveness of gaze modeling; and Global rotation and translation error measures head-pose stability within the shared 3D environment.

**SI for Diversity (SID).** To evaluate behavioral diversity, we apply the SID metric. Motion sequences are clustered in a feature space using k-means with $k = 40$. SID is computed as the entropy of the resulting cluster-assignment histogram, with higher values indicating broader and more varied expressive patterns.

Table 6. Quantitative and human evaluation of LLM-based spatial generation. **tMSE** measures the distance between predicted and groundtruth translations and the lower the better, while **SAS** measures perceptual quality (1-5 scale).

| Method | tMSE ↓ | SAS ↑ |
|---|---|---|
| Zero-Shot | 2.90 | 3.15 |
| $k$-NN | 1.84 | 2.65 |
| **Few-Shot (Ours)** | **0.72** | **4.62** |

## 4. LLM-based Spatial Control Evaluation

To validate the effectiveness of our LLM-based spatial control mechanism, we randomly selected a test set of 1024 videos from our raw conversational video dataset, and use Gemini to annotate each video with a text prompt describing the environment and people's spatial relationship by looking at the first and last video frames. Such spatial annotations range from intimate proximities (e.g., "whispering in ear") to distant interactions (e.g., "shouting across a hall").

We conducted a study comparing our proposed few-shot prompting strategy against (1) a standard zero-shot baseline where the prompt is exactly the same as in Sec. 2.2 but without the three examples, and (2) a $k$-nearest neighbor baseline ($k = 5$) where we retrieval the top-$k$ closest samples from the dataset computing CLIP score and use their average as the output. We generate predicted initial first-frame translation based on the textual annotation.

We evaluate performance using two metrics:

1. **Translation MSE (tMSE):** The mean squared distance between the generated spatial layout and the 3D groundtruth translation obtained by fitting face model to the videos.
2. **Semantic Alignment Score (SAS):** A human evaluation metric ($N = 6$ participants) where users rate the correspondence between the text prompt and 3D layout on a Likert scale from 1 (Poor) to 5 (Excellent).

As shown in Tab. 6, our few-shot strategy significantly outperforms the zero-shot and $k$-NN baselines, both of which often generate plausible but generic layouts. In contrast, our few-shot approach achieves superior semantic alignment, demonstrating that in-context examples are critical for generating spatially meaningful 3D coordinates.

Table 7. Performance comparison on the DualTalk OOD Dataset (384 unseen clips). Our method demonstrates superior generalization on unseen identities compared to the baseline trained specifically on the DualTalk domain. Lower is better for all metrics.

| Method | FD ↓ | P-FD ↓ | vMSE-S ↓ | vMSE-L ↓ |
|---|---|---|---|---|
| DualTalk | 34.12 | 45.20 | 8.95 | 9.12 |
| **Ours** | **14.55** | **31.05** | **4.92** | **6.30** |

## 5. Out-Of-Distribution Dataset Evaluation

To rigorously evaluate the generalization capabilities of our framework, we conducted an additional experiment using the Out-of-Distribution (OOD) validation split of the DualTalk dataset. As described in the DualTalk supplementary material, this specific subset consists of 384 video clips featuring identities and conversation scenarios that are strictly excluded from their training set. This experiment tests a model's ability to handle unseen speakers and novel interactive contexts. We compare our method against the DualTalk baseline by converting all face models into our face model representation to align with the main evaluation table. As shown in Tab. 7, our method significantly outperforms the baseline across all metrics.

While DualTalk is specialized for its specific capture distribution, it struggles to generalize to these unseen OOD identities, resulting in higher geometric error (vMSE) and degraded interaction scores (P-FD). In contrast, our model, having been pre-trained on our massive-scale corpus of 2 million dyadic pairs, exhibits exceptional robustness. This confirms that our large-scale training strategy yields a model that is not only accurate but highly generalizable to diverse, in-the-wild identities.

## 6. Limitations and Future Works

While our approach effectively models dyadic facial interactions, several limitations remain.

**Listener Behavior Category Control.** Regarding listener modeling, our method tends to learn average reactive behaviors, sometimes resulting in generic feedback. Future work could aim to explicitly model discrete listener states (e.g., specific nodding, head shaking, or confusion) to generate more semantically meaningful non-verbal cues.

**Audio Overlapping.** Handling simultaneous speaking with heavy audio overlap remains challenging; imperfect source separation in these scenarios can occasionally lead to degraded lip synchronization and wrong distribution of lip motions to corresponding speakers.

**Personality and Emotion Control.** Our system could benefit from explicit personality control, which might significantly influence avatar behaviors especially in the domain of head movements and eye gaze contact. This could potentially free the users from manually conditioning the animation on specific traits (e.g., "extroverted" or "shy") through controlling the spatial layouts.

**Full Body Generation.** Our scope is currently limited to facial and head dynamics, overlooking the communicative value of hand gestures and body posture. Future work should integrate full-body motion synthesis to create a holistic conversational agent.

**Real-Time Alternatives.** Our reliance on a diffusion architecture necessitates iterative denoising steps during inference. This results in significantly higher computational costs and latency compared to single-pass autoregressive models, presenting a trade-off between generation quality and real-time applicability that could be addressed via distillation techniques.