EchoMask: Speech-Queried Attention-based Mask Modeling for Holistic Co-Speech Motion Generation

Xiangyue Zhang*
Wuhan University
Wuhan, China
xiangyuezhang@whu.edu.cn

Jianfang Li*
Tongyi Lab, Alibaba Group
Hangzhou, China
wuhui.ljf@alibaba-inc.com

Jiaxu Zhang
Wuhan University
Wuhan, China
zjiaxu@whu.edu.cn

Jianqiang Ren
Tongyi Lab, Alibaba Group
Hangzhou, China
jianqiang.rjq@alibaba-inc.com

Liefeng Bo
Tongyi Lab, Alibaba Group
Seattle, USA
liefeng.bo@alibaba-inc.com

Zhigang Tu†
Wuhan University
Wuhan, China
tuzhigang@whu.edu.cn

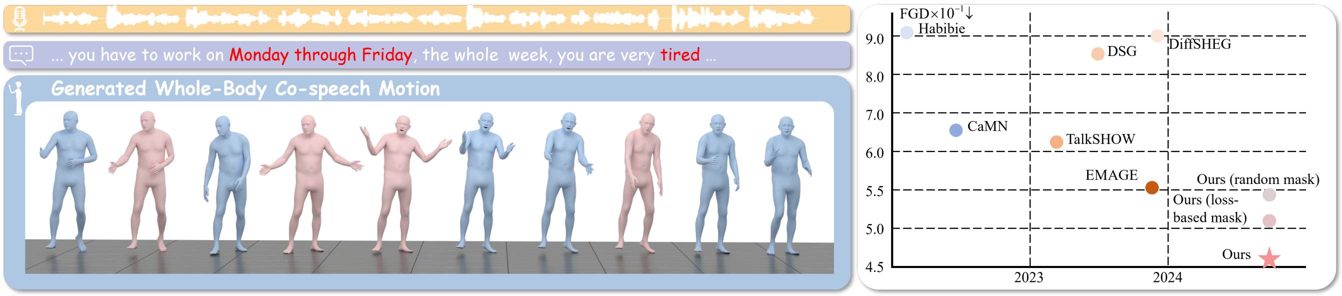


Figure 1: On the left, our EchoMask generates expressive, semantically aligned whole-body co-speech motions from speech input. On the right, our method outperforms previous approaches by leveraging speech-queried attention mask modeling and motion-audio embedding, surpassing both random and loss-based strategies.

Abstract

Masked modeling framework has shown promise in co-speech motion generation. However, it struggles to identify semantically significant frames for effective motion masking. In this work, we propose a speech-queried attention-based mask modeling framework for co-speech motion generation. Our key insight is to leverage motion-aligned speech features to guide the masked motion modeling process, selectively masking rhythm-related and semantically expressive motion frames. Specifically, we first propose a motion-audio alignment module (MAM) to construct a latent motion-audio joint space. In this space, both low-level and high-level speech features are projected, enabling motion-aligned speech representation using learnable speech queries. Then, a speech-queried attention mechanism (SQA) is introduced to compute frame-level attention scores through interactions between motion keys and speech queries, guiding selective masking toward motion frames with high attention scores. Finally, the motion-aligned speech features are also injected into the generation network to facilitate co-speech motion generation. Qualitative and quantitative evaluations confirm that our method outperforms existing state-of-the-art approaches, successfully producing high-quality co-speech motion. Project page: https://xiangyue-zhang.github.io/EchoMask

CCS Concepts

- **Computing methodologies** → *Motion processing*.

Keywords

Co-speech motion generation, speech to gesture generation, mask motion modeling

ACM Reference Format:
Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. 2025. EchoMask: Speech-Queried Attention-based Mask Modeling for Holistic Co-Speech Motion Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754847

1 Introduction

Holistic co-speech motion generation, integrating non-verbal cues such as body poses [13, 18], hand gestures [30, 38], and facial expressions [14, 53] aligned with speech, is an increasingly prominent field in computer vision.

Generating detailed and expressive whole-body motions remains a computationally intensive and technically challenging problem. Masked motion modeling (MMM) [19, 42, 43] has recently gained

*Both authors contributed equally to this research.
†Corresponding Author.

MM '25, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754847

attention as a promising framework to address these challenges [26, 35, 39]. MMM begins by discretizing continuous motion sequences into compact motion tokens via vector quantization (VQ) [46], enabling efficient learning in a discrete latent space. During training, a subset of motion frames or tokens is masked, and the model is tasked with reconstructing the missing elements from the remaining context. A critical factor in the success of MMM lies in the masking strategy—specifically, which frames or tokens are selected for masking. Evidence from masked image modeling (MIM) [6, 21, 52] shows that mask patterns significantly affect model performance by shaping the focus of the learning process [27, 48, 51]. However, current masking approaches in MMM often rely on random or loss-based guided strategies, which struggle to consistently identify semantically important motion segments. This limitation hampers the model's ability to generate expressive, temporally aligned motions.

By delving deeper into the underlying limitations, we identify the core bottleneck in advancing semantically grounded masked motion modeling: the effectiveness of semantically rich motion frames mask selection strategies. Existing methods [19, 26, 35, 39] predominantly adopt random or loss-based masking strategies. Random masking often fails to target semantically meaningful regions due to the sparse distribution of semantic content in motion sequences. Loss-based masking, which prioritizes frames with high reconstruction error, assumes these are semantically rich. However, this assumption does not always hold—high reconstruction loss may simply reflect abrupt yet uninformative transitions. For instance, a sharp change in hand position might signal the end of a sentence rather than a meaningful gesture. As a result, such strategies struggle to accurately identify semantically significant frames, ultimately limiting the quality of speech-conditioned motion generation. Moreover, prior methods mask at the token level using discrete code indices, which lose fine-grained motion details and hinder accurate detection of frames critical for motion intent and speech alignment.

Based on this observation, we raise a central question: Can speech be used as a query to identify semantically important motion frames worth focusing on during masked modeling? To explore this, we propose a new masked motion modeling framework, EchoMask, for co-speech motion generation. Our core motivation is to leverage speech not only as a conditioning signal but also as an informative query that aligns with latent motion representations to pinpoint motion frames with high semantic relevance.

In EchoMask, we introduce a speech-queried, attention-based masked motion modeling framework that leverages motion-aligned speech features to guide the masked motion modeling process, selectively masking semantically rich motion frames. To this end, we propose two key innovations: a hierarchical joint embedding module for motion-audio alignment (MAM) and a speech-queried attention mechanism (SQA) for keyframe selection.

Specifically, MAM is a hierarchical cross-modal alignment module, designed to project paired latent motion and audio inputs into a shared latent space to effectively align the heterogeneous motion and audio modalities. Inspired by CLIP-style dual-tower architectures and [32, 34], MAM utilizes both low-level and high-level HuBERT features, interacting with learnable speech queries via self-attention and cross-attention to generate motion-aligned speech features. The speech queries and latent motion are passed through a shared transformer network, encouraging modality-invariant feature extraction. This design not only reduces parameter redundancy but also promotes more consistent alignment between speech and motion representations, facilitating improved generation fidelity.

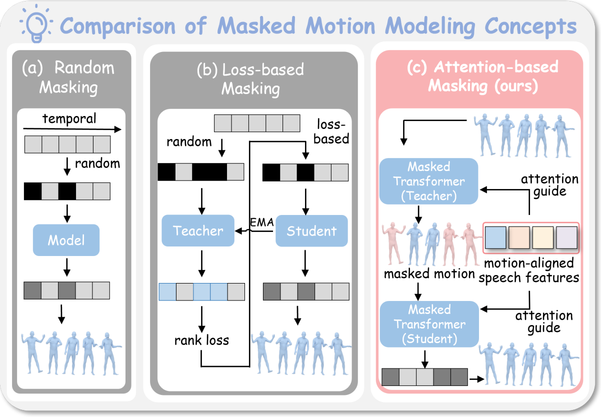

Comparison of Masked Motion Modeling Concepts
(a) Random Masking
temporal
random
Model
(b) Loss-based Masking
random
loss-based
Teacher
EMA
Student
rank loss
(c) Attention-based Masking (ours)
Masked Transformer (Teacher)
attention guide
masked motion
motion-aligned speech features
Masked Transformer (Student)
attention guide

Figure 2: Comparison of masked motion modeling concepts, where three masking strategies are reported: (a) random masking, (b) loss-based masking, and (c) our proposed speech-queried attention-based masking. Unlike the previous approaches, our method selectively highlights motion frames (in red) that are more semantically aligned with the input speech.

SQA then computes frame-level attention scores by modeling the interaction between motion keys and motion-aligned speech features produced by the MAM module. Frames receiving higher attention scores are identified as semantically informative and are preferentially masked. This targeted masking strategy encourages the model to focus on speech-relevant motion patterns during training, as illustrated in Figure 2. By selectively masking keyframes based on attention scores, the model is guided to learn meaningful associations between speech and motion. Finally, the motion-aligned speech features are injected into the generation network, enhancing the quality and synchronization of the synthesized co-speech motion.

We comprehensively evaluate EchoMask on public co-speech motion generation benchmarks. Our qualitative and quantitative results demonstrate that EchoMask significantly outperforms existing methods in terms of semantic alignment, motion realism, and generation diversity.

Our contributions are summarized below:

- We propose EchoMask for co-speech motion generation, a novel masked motion modeling framework that utilizes motion-aligned speech features to mask semantically important motion frames.
- We introduce MAM, a hierarchical cross-modal alignment module that embeds motion and audio into a shared latent space, producing motion-aligned speech features. We also propose SQA, a speech-queried attention mechanism that computes frame-level attention scores, enabling selective identification of semantically important motion frames.
- Extensive experiments demonstrate the superiority of EchoMask over state-of-the-art methods in terms of semantic alignment, motion quality, and generation diversity.

2 Related Work

Holistic Co-speech Motion Generation Holistic co-speech motion generation involves producing coordinated movements of the face, hands, and torso from speech input. Most existing methods focus on isolated parts rather than generating fully integrated body motion. Early rule-based methods [8, 25, 28, 29] mapped speech to gestures using linguistic rules but required extensive manual effort. Recent data-driven approaches use deep generative models—such as GANs [1, 20, 41, 49], VQ-VAEs [3, 37, 44], normalizing flows [45, 55], and diffusion models [2, 11, 22, 54, 60]—to learn complex motion distributions from data.

Habibie et al. [20] introduced a CNN-based model that outputs 3D face, hand, and body motion. Later methods improved realism and synchronization using discrete latent codes. For instance, TalkSHOW [56] synchronizes body and hand gestures with a VQ-VAE. DiffSHEG [12] enables unidirectional information flow between face and body through separate encoders.

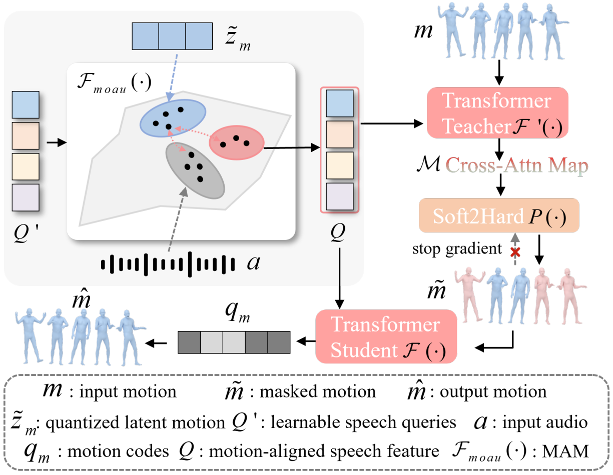
Masked modeling has further improved efficiency. EMAGE [35] and SemTalk [59] use random motion masking, while ProbTalk [39] leverages PQ-VAE [50] to reconstruct masked inputs. However, these methods rely on random or causal masking, which limits the model's ability to focus on semantically important motion frames. In our work, we propose a speech-queried attention-based mask modeling that utilizes motion-aligned speech features to identify semantically important motion frames for co-speech motion generation.

Masking Strategies in Masked Modeling Masking plays a key role in masked modeling by influencing what the model learns. Existing strategies fall into three categories: (i) **Random masking**. Used in early models like BERT [15] and extended to vision in ViT [16], with refinements in BEiT [6]. Generative models such as MaskGIT [10] and Muse [9] also apply random masking for representation learning. (ii) **Loss-based masking**. AdaMAE [5] assumes that semantically meaningful patches are harder to reconstruct and therefore prioritizes masking regions with high reconstruction loss to guide the model toward learning informative content. HPM [48] focuses on increasing task difficulty by selecting patches the model finds hard to reconstruct, using a masking strategy that progresses from easy to hard examples. (iii) **Attention-based masking**. AMT [40], and AttMask [27] use class tokens to mask semantically rich regions, while MILAN [23] applies CLIP-based knowledge distillation for high-importance patch selection. While effective, these methods struggle to flexibly identify semantically important content guided by external signals. Inspired by [51], we introduce speech-queried attention masking that computes frame-level attention scores by evaluating the interaction between motion keys and motion-aligned speech features to identify significant frames.

3 Method

3.1 Preliminary on Masked Motion Modeling

Masked motion modeling (MMM) draws inspiration from masked image modeling [5, 9, 10, 48] and aims to learn expressive motion representations by reconstructing missing portions of a motion sequence based on visible frames or tokens. Given a motion sequence consisting of N frames, we denote it as a set of frame-wise units $M = \{\mathbf{m}_i\}_{i=1}^{N}$, where each $\mathbf{m}_i$ represents either joint positions, rotation representations, or quantized latent tokens at frame i.



$\tilde{z}_m$
$\mathcal{F}_{moau}(\cdot)$
Q'
a
Q
m
Transformer Teacher $\mathcal{F}'(\cdot)$
$\mathcal{M}$ Cross-Attn Map
Soft2Hard $P(\cdot)$
stop gradient
$\tilde{m}$
$\hat{m}$
q_m
Transformer Student $\mathcal{F}(\cdot)$
m : input motion
$\tilde{m}$: masked motion
$\hat{m}$: output motion
$\tilde{z}_m$: quantized latent motion
Q' : learnable speech queries
a : input audio
q_m : motion codes
Q : motion-aligned speech feature
$\mathcal{F}_{moau}(\cdot)$: MAM

Figure 3: An overview of the EchoMask pipeline. Given audio a and quantized latent motion $\tilde{z}_m$, MAM $\mathcal{F}_{\text{moau}}$ produces motion-aligned speech features Q in a shared latent space. The Transformer Teacher $\mathcal{F}'$ computes a cross-attention map $\mathcal{M}$ between Q and input motion m, which guides the Soft2Hard module $P(\cdot)$ to mask semantically important frames. The Transformer Student $\mathcal{F}$ then generates motion codes q_m from the masked motion $\tilde{m}$ and Q, and the output motion $\hat{m}$ is decoded via the RVQ-VAE decoder.

A masking ratio $r \in (0, 1)$ is applied to randomly select a subset of frames or tokens $\mathcal{S} \subset \{1, \ldots, N\}$, such that $|\mathcal{S}| = \lfloor rN \rfloor$. The selected elements are masked and replaced with learnable mask frames or tokens. The remaining unmasked subset $S_r = \{\mathbf{m}_i : i \notin \mathcal{S}\}$ is passed through an encoder to extract contextual motion features. The model is then trained to reconstruct the masked content using both the encoded visible frames and the learnable mask representations. A decoder receives the encoded visible context and the mask embeddings to generate predictions for the masked frames or tokens. The training objective is to minimize the reconstruction error between the predicted and ground-truth motion values for the masked subset:

$$\mathcal{L}_{\text{MMM}} = \sum_{i \in \mathcal{S}} \left\| \mathbf{m}_i^d - \mathbf{m}_i \right\|_2^2 . \tag{1}$$

where $\mathbf{m}_i^d$ denotes the reconstructed output of the decoder for the i-th masked frame or token. Through this process, the model is encouraged to learn the underlying spatiotemporal structure and semantic consistency of human motion. MMM thus provides a powerful pretext task for learning generalizable and context-aware motion representations.

3.2 Overview

As illustrated in Figure 3, EchoMask consists of two core modules that work together to enable semantically grounded co-speech motion generation: (1) a hierarchical motion-audio alignment module (MAM) for generating motion-aligned speech features, and (2) a speech-queried attention mechanism (SQA) for identifying and masking key motion frames.

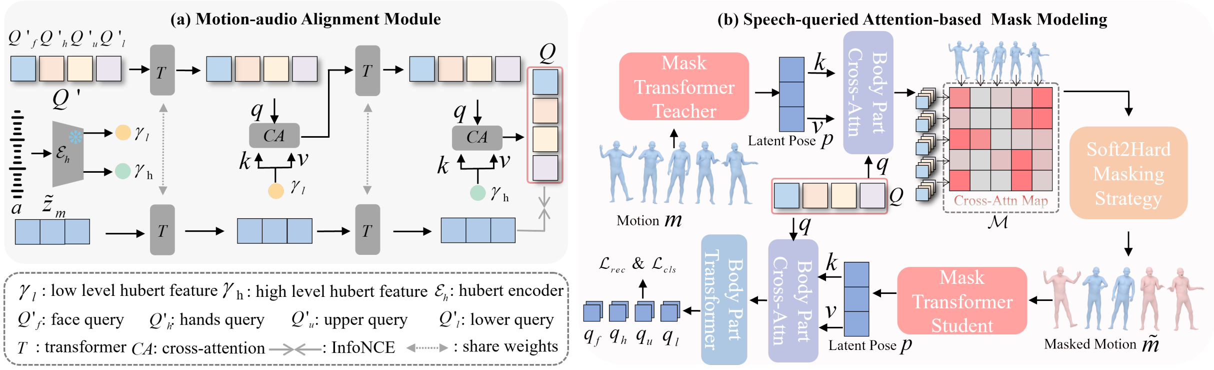

(a) Motion-audio Alignment Module
Q'_f Q'_h Q'_u Q'_l
Q'
T
q
CA
k
v
γ_l
γ_h
ℰ_h
a
z̃_m
Q
γ_l : low level hubert feature γ_h : high level hubert feature ℰ_h: hubert encoder
Q'_f: face query Q'_h: hands query Q'_u: upper query Q'_l: lower query
T : transformer CA: cross-attention : InfoNCE : share weights
(b) Speech-queried Attention-based Mask Modeling
Mask Transformer Teacher
k
v
Latent Pose p
Body Part Cross-Attn
q
Q
Cross-Attn Map
ℳ
Soft2Hard Masking Strategy
Motion m
ℒ_rec & ℒ_cls
q_f q_h q_u q_l
Body Part Transformer
Body Part Cross-Attn
Mask Transformer Student
Masked Motion m̃

Figure 4: Architecture of EchoMask. (a) MAM projects motion and audio into a shared latent space. Learnable speech queries Q' are refined through hierarchical cross-attention with HuBERT features (γ_l, γ_h) and jointly processed with quantized latent motion $\tilde{z}_m$ via a shared transformer, optimized with contrastive loss. (b) Given m, mask transformer teacher computes a cross-attention map $\mathcal{M}$ between latent poses p and motion-aligned speech features Q, identifying semantically important frames. These frames are masked via a Soft2Hard strategy to produce $\tilde{m}$, which the student transformer uses to generate motion tokens.

Given input audio a and quantized latent motion tokens z_m, MAM $\mathcal{F}_{\text{moau}}$ first projects paired motion and speech features into a shared latent space by leveraging both low-level and high-level HuBERT features. Learnable speech queries interact with these features via cross-attention and self-attention, resulting in motion-aligned speech representations Q:

$$\mathcal{F}_{\text{moau}} : (a, z_m, Q'; \theta_{\text{moau}}) \mapsto Q, \tag{2}$$

where θ_{moau} denotes MAM's parameters and Q is used to guide the masking process and motion generation.

To identify semantically important frames, a Transformer Teacher $\mathcal{F}'$ (an EMA copy of the Student $\mathcal{F}$) computes a cross-attention map $\mathcal{M}$ between motion input m and speech features Q.

The $\mathcal{M}$ is then passed to the Soft2Hard masking module $P(\cdot)$, which masks the frames with high attention scores to yield the masked motion sequence $\tilde{m}$:

$$\tilde{m} = P(m, \mathcal{M}). \tag{3}$$

The fused representation q_m, obtained from $\mathcal{F}$, Q and $\tilde{m}$, is then decoded into the final motion $\hat{m}$ using an RVQ-VAE decoder [7, 19, 58]. To further disentangle motion semantics across the body, we adopt a part-wise decoder structure following [4, 35], dividing the body into face, hands, upper body, and lower body segments.

3.3 Motion-Audio Joint Embedding Learning

Prior works [11, 12, 35, 56] condition on static speech features without addressing the modality gap between audio and motion, leading to weak alignment and degraded generation quality. We introduce a Motion-Audio Alignment Module (MAM, Figure 4a), which maps both modalities into a shared latent space through two stages: Hierarchical Speech Query Encoding and Latent Space Alignment.

Hierarchical Speech Query Encoding. MAM first focuses on refining learnable speech queries through hierarchical feature fusion. We extract both low-level and high-level audio features from a pretrained HuBERT model [24], denoted as γ_l and γ_h, respectively. These features provide complementary information—phonetic timing from shallow layers and semantic context from deeper layers. Learnable speech queries Q' are initialized randomly and progressively refined by interacting with these hierarchical speech features through cross-attention layers. The first cross-attention layer fuses local acoustic details, while the second contextualizes them using global semantics. The output Q' serves as the refined set of speech queries for motion alignment.

Latent Space Alignment. After encoding, the refined speech queries Q' are aligned with the quantized latent motion sequence $\tilde{z}_m$ using a shared Transformer $T(\cdot)$. Both streams are passed through the same transformer layers to ensure feature consistency under identical inductive biases. This shared backbone allows the queries and motion tokens to be co-trained and updated via synchronized gradients. Formally:

$$\hat{Q}, \hat{z}_m = T(Q'), T(\tilde{z}_m), \tag{4}$$

where $\hat{Q}$ and $\hat{z}_m$ denote the final embeddings of speech queries and motion tokens, respectively.

To enforce tight alignment between the modalities, we introduce a contrastive loss based on InfoNCE at both the frame level and the global (sentence) level. For a batch of B paired sequences, we define:

$$\mathcal{L}_{\text{align}} = -\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(\hat{q}_i, \hat{z}_i)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(\hat{q}_i, \hat{z}_j)/\tau\right)}. \tag{5}$$

where $\hat{q}_i$ and $\hat{z}_i$ are either the frame-level or pooled sentence-level embeddings for the i-th pair, $\text{sim}(\cdot,\cdot)$ is a similarity function (e.g., cosine similarity), and τ is a temperature hyperparameter. This contrastive objective ensures that motion and speech embeddings corresponding to the same input are drawn closer together, while embeddings from different pairs are pushed apart.

3.4 Speech-queried Attention Mechanism

Although random masking assumes uniform semantic distribution across motion sequences, this rarely holds—semantic cues are often concentrated in frames aligned with speech. Loss-based masking [26] misidentifies semantic content by focusing on high-error or dense frames. To address this, we propose a Speech-Queried Attention (SQA) mechanism (Figure 4b), which leverages cross-modal attention between speech and motion to identify and mask semantically aligned frames.

Speech-queried Cross-Attention. Given the motion-aligned speech features Q from the MAM and an input motion sequence m, we first obtain its latent pose representation p. The body-part cross-attention module then computes a cross-attention map $\mathcal{M}$ using p as keys and Q as query, processed by a mask transformer teacher. Since the motion is factorized into four regions—face, hands, upper body, and lower body—the speech queries are also projected into a part-aware latent space, enabling alignment between each body part and the speech. The resulting attention map, defined as $\mathcal{M} = \sum \mathcal{M}_{\text{parts}}$, where $\mathcal{M}_{\text{parts}} \in \{\mathcal{M}_{\text{face}}, \mathcal{M}_{\text{hands}}, \mathcal{M}_{\text{upper}}, \mathcal{M}_{\text{lower}}\}$, captures fine-grained semantic relevance between speech and motion. This facilitates precise identification of which motion frames and parts are most aligned with the speech content. To derive a frame-level importance score $s \in \mathbb{R}^T$, we aggregate attention scores across all motion-aligned speech features Q:

$$s_j = \sum_{i=1}^{T} \mathcal{M}_{i,j}, \quad j \in \{1, \ldots, T\}, \tag{6}$$

where s_j represents the semantic significance of the j-th motion frame with respect to the motion-aligned speech features. Frames with higher s_j values are considered semantically richer and more relevant to speech dynamics.

To promote alignment with semantically meaningful motion, we apply soft supervision $\mathcal{L}_{sem}$to the Student's frame-level attention scores using binary cross-entropy loss against soft or binary semantic labels. This auxiliary objective guides the Student to focus on important frames, and through EMA updates, the Teacher gradually inherits this behavior, yielding more interpretable and speech-consistent attention maps for guiding masking.

Soft-to-Hard Masking Strategy. To avoid prematurely masking high semantic frames, which can hinder effective learning in the early stages, we employ a *soft-to-hard* masking strategy that aligns with the model's learning progression. The `argsort`-based masking, which deterministically selects the top semantic frames, is treated as a "hard" masking strategy. In contrast, the "soft" variant samples frames based on a probability distribution defined by s_j, allowing for stochastic selection.

During training, we gradually shift from soft to hard masking by adjusting their proportions over epochs. Specifically, at training epoch t, the soft and hard mask ratios are updated as:

$$\begin{aligned} \alpha_t^s &= \alpha_0^s + \frac{t}{T}(\alpha_T^s - \alpha_0^s), \\ \alpha_t^h &= \alpha_0^h - \frac{t}{T}(\alpha_0^h - \alpha_T^h), \\ \alpha_t^r &= \alpha - \alpha_t^s - \alpha_t^h. \end{aligned} \tag{7}$$

where α_t^s, α_t^h, and α_t^r denote the proportions of soft, hard, and random masks at epoch t, and $\alpha_0^s, \alpha_T^s, \alpha_0^h, \alpha_T^h$ are predefined initial and final values. α is mask ratio.

3.5 Inference

As shown in Figure 5, EchoMask generates co-speech motion using speech input alone. Unlike prior methods that directly use features from a pretrained audio encoder as static conditions, we leverage motion-aligned speech representations to guide generation. Specifically, hierarchical speech features are first extracted from a pretrained HuBERT encoder $\mathcal{E}_h$, then fused with learnable speech queries via cross-attention in the MAM. This results in refined, motion-aligned speech features Q, which capture both low-level prosody and high-level semantic intent aligned with the motion.

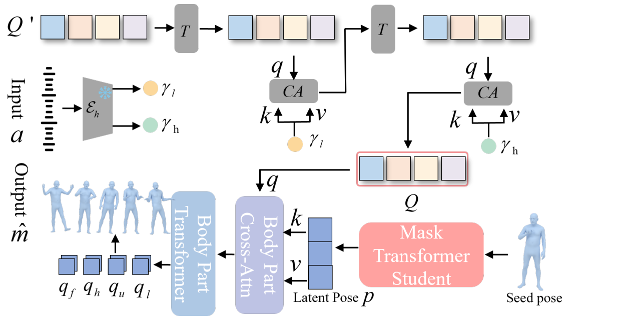


Figure 5: Pipeline of Inference. EchoMask takes speech as the sole input and employs hierarchical HuBERT features to guide a mask transformer, initialized with four seed pose frames. The model predicts motion tokens with the guidance of motion-aligned speech features Q that are subsequently decoded by the RVQ-VAE decoder to generate whole-body motion.

Using only four seed motion frames and the aligned speech features Q, we initialize the masked motion sequence and input it to the transformer student. The model predicts the full sequence of motion tokens, which are then decoded by the RVQ-VAE decoder to generate the final co-speech motion.

4 Experiments

4.1 Experimental Setup

Datasets. For training and evaluation, we use the BEAT2 dataset [35], which contains 60 hours of high-quality finger motion data from 25 speakers (12 female, 13 male). The dataset includes 1762 sequences, each averaging 65.66 seconds, where speakers respond to daily inquiries. We divide the dataset into training (85%), validation (7.5%), and test (7.5%) sets. To ensure a fair comparison, we follow [35] and use data from Speaker 2 for training and validation.

Implementation Details. Our model is trained on a single NVIDIA A100 GPU for 200 epochs with a batch size of 64. The RVQ-VAE is downscaled by 4. The residual quantization has 6 layers, a codebook size of 256, and a dropout rate of 0.2. The training uses the ADAM optimizer with a 1e-4 learning rate. During inference, we use a four-frame seed pose to initialize each motion clip. For consecutive clips, the last four frames of the previous segment are overlapped and reused as the seed for the next, following [35].

Metrics. We assess the quality of generated body gestures using the FGD metric [57], which evaluates how closely the distribution of generated gestures aligns with ground truth (GT), providing a measure of realism. Gesture diversity is quantified using DIV [30], calculated as the average L1 distance across multiple gesture clips to capture motion variation. To evaluate speech-motion synchrony, we employ BC [31], which measures the temporal alignment between gesture rhythm and audio beats. For facial expression accuracy, we use two reconstruction metrics: vertex MSE [54] to assess positional

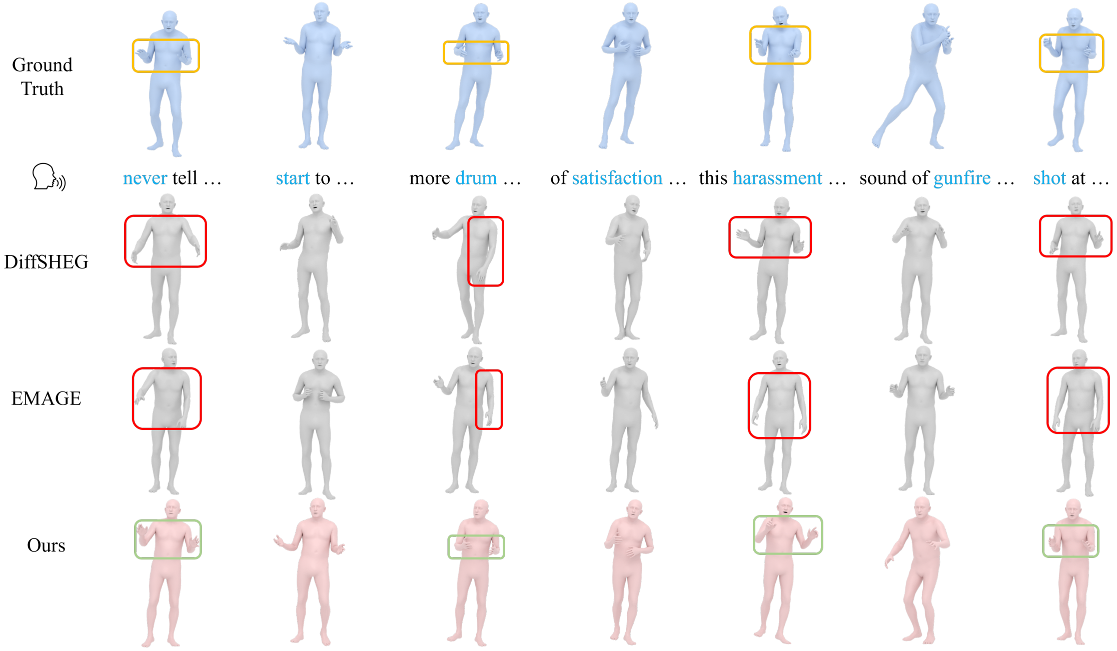


Figure 6: Visual comparison. Red boxes highlight implausible or uncoordinated motions, while green boxes indicate coherent and semantically appropriate results. Our EchoMask consistently generates co-speech motions that are semantically aligned with ground truth. For instance, when articulating "never" and "start", our model positions both hands in a poised gesture near the torso, reflecting a thoughtful and intentional motion, whereas prior methods such as DiffSHEG and EMAGE either generate imbalanced hand postures or ambiguous limb placements.

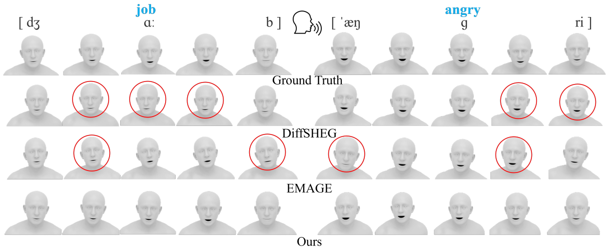


Figure 7: Facial Comparison.

differences and vertex L1 difference (LVD) [56] to quantify discrepancies between GT and generated facial vertices. Additionally, we conduct a user study to provide a more comprehensive evaluation.

4.2 Qualitative Results

Qualitative Comparisons. As illustrated in Figure 6, our EchoMask consistently produces co-speech motions that are both semantically aligned with the ground truth and physically expressive. In contrast, baseline methods such as DiffSHEG and EMAGE often generate gestures that are misaligned with the underlying speech semantics and visually implausible.

In the case of "drum", EchoMask captures the rhythmic semantics by swinging one arm outward in a dynamic arc, a detail that other methods miss—often producing static or downward-pointing arms. Similarly, for the term "satisfaction", our method aligns the gesture with the meaning by raising the right arm close to the chest in a self-reflective manner, while other baselines lack such subtlety. Interestingly, when expressing "harassment" and "gunfire", EchoMask emphasizes the tension through body posture—one arm bent with visible muscular contraction and the upper body leaning slightly forward—accurately conveying urgency or defensive response. Competing DiffSHEG and EMAGE display either stiffness or lack of spatial coordination in these challenging cases.

For the word "shot," our model conveys the implied action with a firm, forward-facing hand posture, where both arms adopt assertive positions—capturing context missed by DiffSHEG and EMAGE, which often produce low or passive hand gestures. EchoMask consistently preserves hand dominance, articulation range, and delivers a richer set of motions aligned with the acoustic-semantic cues of speech.

This qualitative evidence underscores the strength of our speech-queried attention mask modeling in driving expressive and semantically grounded motion synthesis.

Figure 7 highlights expression errors in baselines, often showing stiff or mistimed facial motion. Our method produces smoother, phoneme-aligned transitions, accurately capturing key articulations like lip closure for [æŋ] and jaw opening for [dʒ], demonstrating the strength of our speech-aligned, part-aware facial synthesis.

Masking Strategy. As illustrated in Figure 8, the random masking strategy (top row) produces a uniform, indiscriminate pattern, often masking low-information frames or temporally scattered segments. The loss-based masking strategy (middle row), which selects frames with high reconstruction error, similarly exhibits limitations. While these frames often reflect abrupt pose transitions or motion discontinuities, they do not reliably correspond to semantically salient content—frequently capturing transitional noise or speech pauses rather than meaningful gestures.

In contrast, our speech-queried masking strategy (bottom row) yields a targeted and speech-synchronized masking pattern. The

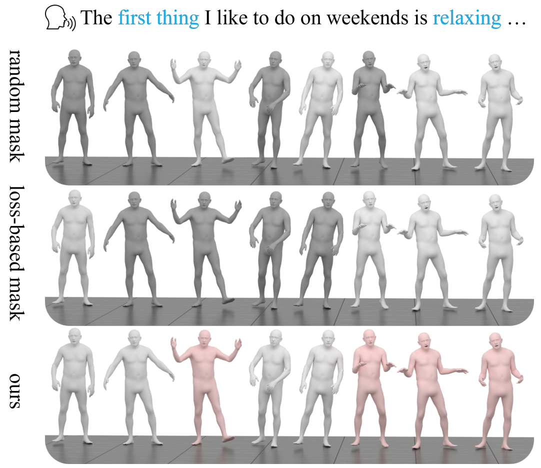


Figure 8: Visualization of generated masked motion by random mask, loss-based mask, and our method. The darker and red motion frames represent those that are masked out.

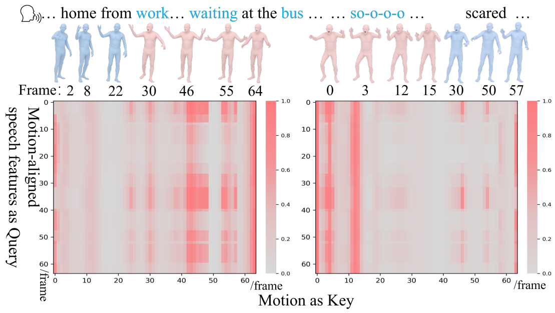
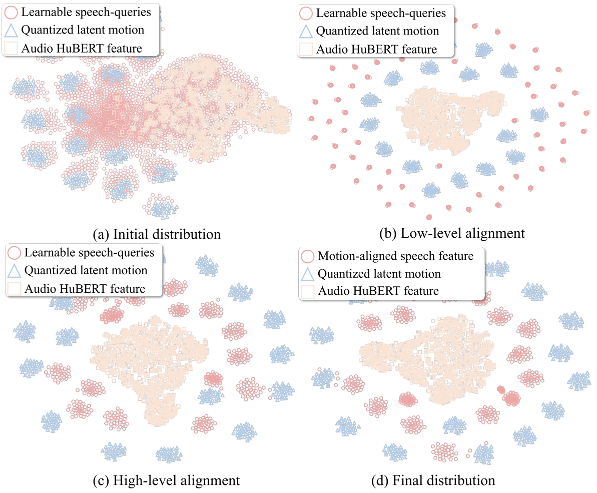


Figure 9: Visualization of the cross-attention map between speech queries and motion frames. Darker red regions indicate higher attention scores. Our model attends to semantically rich motion frames that align closely with key speech tokens such as "work," "waiting," and "so-o-o-o," demonstrating effective cross-modal alignment.

masked frames are concentrated around semantically rich regions, closely aligned with gesture peaks that convey the intent of the spoken words—for example, the emphatic upward motion for "first" and the fluid, relaxed hand movement for "relaxing." By explicitly grounding the masking in cross-modal attention, our approach ensures the model learns to reconstruct motion segments that are both contextually and semantically significant.

Cross-Attn Map. As shown in Figure 9, our model assigns high attention scores to semantically expressive regions, such as "work," "waiting," and the elongated syllables in "so-o-o-o," demonstrating its ability to capture both word-level semantics and prosodic emphasis. Unlike previous approaches with diffuse or noisy attention, our speech-queried mechanism produces focused and interpretable patterns. The attention selectively highlights gesture-relevant frames while de-emphasizing idle motions, enabling more effective masking and improving alignment between speech and motion.

The Effectiveness of MAM. Figure 10 illustrates the progressive refinement of speech features through our MAM module. Each point represents a frame-level feature. Initially (top-left), the learnable speech queries (red circles) are scattered and poorly aligned with motion tokens (blue triangles) and audio features (peach squares),



(a) Initial distribution (b) Low-level alignment

(c) High-level alignment (d) Final distribution

Figure 10: t-SNE [47] visualization of the shared latent space. After MAM, motion-aligned speech features (red circles), quantized latent motion tokens (blue triangles), and HuBERT audio features (peach squares) form well-aligned and semantically coherent clusters, illustrating effective modality fusion and cross-modal alignment.

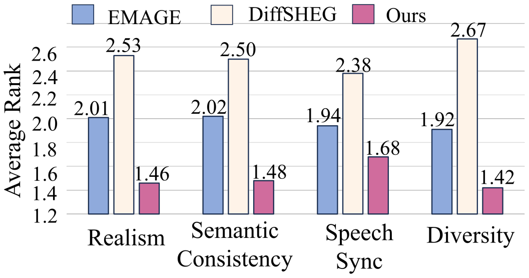


Figure 11: Results of the user study.

indicating weak semantic grounding. After incorporating low-level hierarchical (top-right) and low-level γ_h(bottom-left) HuBERT features γ_l, the distribution becomes more structured. In the final stage (bottom-right), the motion-aligned speech features form compact clusters that are well integrated between audio and motion, demonstrating successful alignment in the shared latent space. This alignment improves attention computation and supports semantically coherent and expressive motion generation, validating the effectiveness of each MAM component.

User Study. To assess perceptual quality, we conducted a user study involving 28 participants with varied backgrounds, who evaluated 10 videos across multiple methods. The evaluation focused on four aspects: realism, alignment with speech semantics, temporal synchrony between motion and speech, and diversity of motion. As illustrated in 11, our method was consistently rated highest.

4.3 Quantitative Results

Comparison with Baselines. Table 2 presents a comprehensive quantitative comparison between EchoMask and a wide range of state-of-the-art methods across facial, non-facial, and holistic motion generation tasks. Our method consistently achieves the best

Method	FGD↓	BC↑	DIV↑	MSE↓	LVD↓
Facial Generation					
FaceFormer [17]	-	-	-	7.787	7.593
CodeTalker [53]	-	-	-	8.026	7.766
Non-facial Gesture Generation					
DisCo [33]	9.680	6.441	9.892	-	-
HA2G [38]	12.14	6.711	8.916	-	-
CaMN [36]	6.644	6.769	10.86	-	-
LivelySpeaker [20]	11.80	6.659	11.28	-	-
DSG [54]	8.811	7.241	11.49	-	-
Holistic Motion Generation					
Habibie *et al.* [20]	9.040	7.716	8.213	8.614	8.043
TalkSHOW [56]	6.209	6.947	**13.47**	7.791	7.771
EMAGE [35]	5.512	7.724	13.06	7.680	7.556
DiffSHEG [12]	8.986	7.142	11.91	7.665	8.673
EchoMask (Ours)	**4.623**	**7.738**	13.37	**6.761**	**7.290**

Table 1: Quantitative comparison with SOTA. Lower values indicate better performance for FMD, FGD, MSE, and LVD, while higher values are better for BC and DIV. For clarity, we report FGD $\times 10^{-1}$, BC $\times 10^{-1}$, MSE $\times 10^{-8}$, and LVD $\times 10^{-5}$. Best results are shown in bold.

performance across almost all metrics. In holistic motion generation, EchoMask outperforms all baselines with the lowest FGD, MSE, and LVD, indicating superior realism, motion accuracy, and temporal smoothness. It also achieves the highest BC, reflecting better rhythm alignment with speech. While TalkSHOW attains the highest DIV, EchoMask remains highly competitive, suggesting it captures a broad range of expressive motions without sacrificing structure or semantic fidelity. In facial generation, EchoMask surpasses strong baselines such as FaceFormer and CodeTalker with significant reductions in MSE and LVD, demonstrating more accurate and emotionally coherent facial articulation.

Ablation Study on Components. As Table 2 shows, replacing the VQ-VAE with the residual variant (RVQ-VAE) results in consistent improvements across metrics. Incorporating our speech-queried attention mask modeling (SQA) leads to further gains. While random and loss-based masking offer moderate improvements, the attention-based approach achieves the lowest FGD and highest motion diversity, highlighting its ability to identify semantically informative frames aligned with the speech content. We also examine the impact of the MAM design. Using either low-level or high-level HuBERT features in isolation yields competitive results; however, fusing both levels significantly enhances motion diversity and beat consistency. Moreover, excluding the alignment loss $\mathcal{L}_{\text{align}}$ causes a clear degradation in both FGD and DIV, underscoring its role in maintaining cross-modal consistency. The complete MAM configuration delivers the most balanced performance across all metrics. Finally, the full EchoMask model outperforms all ablated variants, validating the effectiveness of the proposed SQA and MAM.

Soft-to-hard Masking Strategy. Figure 12 reports an ablation study examining the impact of various soft-to-hard masking schedules on generation quality, evaluated by FGD. The results indicate that a balanced configuration ($\alpha = 0.5$) with a smooth transition from soft to hard masking ($\alpha_0^h = 0, \alpha_T^h = 0.3; \alpha_0^s = 0.3, \alpha_T^s = 0$) yields the best performance. This highlights the advantage of starting with probabilistic frame sampling, which encourages broader motion exploration, and progressively shifting to deterministic selection based on attention, allowing the model to focus on semantically salient frames as training advances.

Method	FGD↓	BC↑	DIV↑	MSE↓	LVD↓
EchoMask(VQ-VAE)	6.664	7.464	10.86	7.225	7.693
+ RVQ-VAE	6.106	7.654	11.68	7.014	7.484
+ Speech-queried Attention Mechanism					
SQA (random mask)	5.889	7.613	12.26	7.122	7.473
SQA (loss-based mask)	5.745	7.607	12.66	7.056	7.454
SQA (ours)	5.455	7.615	12.83	6.976	7.364
+ Hierarchical Motion-audio Alignment					
MAM (low-level feature)	5.679	7.647	13.15	7.153	7.478
MAM (high-level feature)	5.442	7.692	13.02	6.903	7.359
MAM (low&high-level feature)	5.420	7.684	13.21	6.977	7.346
MAM ($\mathcal{L}_{align}$)	5.887	7.571	12.78	6.988	7.393
MAM (ours)	5.029	7.690	13.32	6.827	7.310
EchoMask (Ours)	**4.623**	**7.738**	**13.37**	**6.761**	**7.290**

Table 2: Ablation study evaluating the effectiveness of each component within the EchoMask.

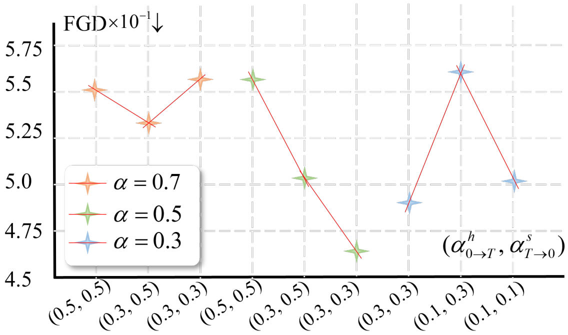


Figure 12: Ablation study on the soft-to-hard masking strategy. We analyze the impact of varying the masking ratios: α_0^h, $\alpha_{0\to T}^h$ for hard masks, and α_0^s, $\alpha_{T\to 0}^s$ for soft masks.

5 Conclusion

In this work, we propose EchoMask, a new masked motion modeling framework for holistic co-speech motion generation. EchoMask identifies semantically expressive co-speech motions using motion-aligned speech features, facilitating effective mask modeling in training. We introduce two key components in EchoMask: a motion-audio alignment module and a speech-queried attention mechanism. The former is a hierarchical cross-modal alignment module that embeds motion and audio into a unified latent space through shared attention and contrastive learning. The latter utilizes motion-aligned speech queries to dynamically identify and mask semantically rich motion frames, improving the model's ability to learn meaningful speech-conditioned motion patterns. Through extensive experiments, EchoMask demonstrates state-of-the-art performance in facial, gestural, and whole-body motion generation. The qualitative results further show its capacity to generate temporally coherent, diverse, and speech-synchronized motions.

Acknowledgments. This work was supported by Alibaba Research Intern Program, the Young Scientists Fund of the National Natural Science Foundation of China No. 624B2110, the National Key Research and Development Program of China No. 2024YFC3015600, the Fundamental Research Funds for Central Universities No.2042023KF0180 & No.2042025KF0053. The numerical calculation is supported by super-computing system in Supercomputing Center of Wuhan University and Tongyi Lab, Alibaba Group.

References

[1] Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. 2022. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20566–20576.

[2] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–20.

[3] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–19.

[4] Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–18.

[5] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. 2023. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14507–14517.

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

[7] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing* 31 (2023), 2523–2533.

[8] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 477–486.

[9] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704* (2023).

[10] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11315–11325.

[11] Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. 2024. Enabling synergistic full-body control in prompt-based co-speech motion generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6774–6783.

[12] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. 2024. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7352–7361.

[13] Kiran Chhatre, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J Black, Timo Bolkart, et al. 2024. Emotional speech-driven 3d body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1942–1953.

[14] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. 2023. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*. 1–13.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[17] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18770–18780.

[18] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3497–3506.

[19] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1900–1910.

[20] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 101–108.

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[22] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. 2024. Co-speech gesture video generation via motion-decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2263–2273.

[23] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. 2022. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049* (2022).

[24] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021), 3451–3460.

[25] Chien-Ming Huang and Bilge Mutlu. 2012. Robot behavior toolkit: generating effective social behaviors for robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 25–32.

[26] Minjae Jeong, Yechan Hwang, Jaejin Lee, Sungyoon Jung, and Won Hwa Kim. [n. d.]. HGM^3: Hierarchical Generative Masked Motion Modeling with Hard Token Mining. In *The Thirteenth International Conference on Learning Representations*.

[27] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. 2022. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*. Springer, 300–318.

[28] Michael Kipp. Universal-Publishers, 2005. Gesture generation by imitation: From human behavior to computer character animation.

[29] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6*. Springer, 205–217.

[30] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11293–11302.

[31] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13401–13412.

[32] Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T Yang. 2024. LaMP: Language-Motion Pretraining for Motion Generation, Retrieval, and Captioning. *arXiv preprint arXiv:2410.07093* (2024).

[33] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM international conference on multimedia*. 3764–3773.

[34] Haiyang Liu, Xingchao Yang, Tomoya Akiyama, Yuantian Huang, Qiaoge Li, Shigeru Kuriyama, and Takafumi Taketomi. 2024. Tango: Co-speech gesture video reenactment with hierarchical audio motion embedding and diffusion interpolation. *arXiv preprint arXiv:2410.04221* (2024).

[35] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2024. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1144–1154.

[36] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*. Springer, 612–630.

[37] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. 2022. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems* 35 (2022), 21386–21399.

[38] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10462–10472.

[39] Yifei Liu, Qiong Cao, Yandong Wen, Huaiguang Jiang, and Changxing Ding. 2024. Towards variable and coordinated holistic co-speech motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1566–1576.

[40] Zhengqi Liu, Jie Gui, and Hao Luo. 2023. Good helper is around you: Attention-driven masked image modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1799–1807.

[41] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. 2021. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10743–10752.

[42] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. 2024. BAMM: bidirectional autoregressive motion model. In *European Conference on Computer Vision*. Springer, 172–190.

[43] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. 2024. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1546–1555.

[44] Kai Shu, Haoyi Zhang, Wai Seng Cheang, Haoyang Wang, Jiechao Gao, et al. 2024. Eggesture: Entropy-guided vector quantized variational autoencoder for co-speech gesture generation. In *ACM Multimedia 2024*.

[45] Shuai Tan, Bin Ji, and Ye Pan. 2024. Flowvqtalker: High-quality emotional talking face generation through normalizing flow and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26317–26327.

[46] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[47] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[48] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. 2023. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10375–10385.

[49] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. 2020. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9332–9341.

[50] Hanwei Wu and Markus Flierl. 2019. Learning product codebooks using vector-quantized autoencoders for image retrieval. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 1–5.

[51] Gongli Xi, Ye Tian, Mengyu Yang, Lanshan Zhang, Xirong Que, and Wendong Wang. 2024. Global Patch-wise Attention is Masterful Facilitator for Masked Image Modeling. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 1751–1760. doi:10.1145/3664647.3681321

[52] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9653–9663.

[53] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12780–12790.

[54] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919* (2023).

[55] Sheng Ye, Yu-Hui Wen, Yanan Sun, Ying He, Ziyang Zhang, Yaoyuan Wang, Weihua He, and Yong-Jin Liu. 2022. Audio-driven stylized gesture generation with flow-based model. In *European Conference on Computer Vision*. Springer, 712–728.

[56] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 469–480.

[57] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

[58] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), 495–507.

[59] Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Ziqiang Dang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. 2024. SemTalk: Holistic Co-speech Motion Generation with Frame-level Semantic Emphasis. *arXiv preprint arXiv:2412.16563* (2024).

[60] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10544–10553.