DISENTANGLED REPRESENTATIONS FOR ZERO-SHOT BEHAVIORAL STYLE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Human behavior is exhibited through bodily motion that not only achieves its intent-driven or communicative goals, but also demonstrates stylistic and idiosyncratic patterns. Existing approaches for generative motion synthesis learn to transfer these styles to novel motions by learning adapters or designing task-specific guidance objectives. However, disentangling style and content within the core motion representation remains an open problem. We present DISCODEC, a general framework for building disentangled representations of style and content in human motion, which can be used for zero-shot style transfer at the decoding stage. Our approach learns these representations through a neural motion codec trained with a Variational Information Bottleneck and SNR-adaptive Transform Quantization to disentangle style from the motion's action content. Two streams of encoders decompose any reference motion into style and content latents, which are then combined to reconstruct the motion, while also ensuring that these latents represent mutually distinct information about the motion. Therefore, any generative motion synthesis task learned using the proposed representation not only inherits zero-shot style transfer capability but can also sample styles from the learned style subspace. We benchmark the style transfer performance against existing approaches and independently demonstrate applications of DISCODEC representation in text-to-motion synthesis and co-speech gesture generation tasks. Reader is urged to watch accompanying video demo.

1 INTRODUCTION

Motion is a fundamental trait of our existence. Interestingly, human motion is quite peculiar in that the same intended motioncan be executed in very different ways depending on the personal style and idiosyncratic behavior of different individuals. These style-driven traits manifest profoundly in how we move and gesticulate, while also forming distinct yet consistent motifs across different goal-specific motions. Ranging from instruction-driven motions to co-speech gestures, stylistic patterns encode *"how"* motion is performed while action-specific cores defines *"what"* it contains or aims to do. However, both style and action exist in a heavily entangled space, with several nuance-factors influencing the way we move. For animation-specific applications, this entanglement makes it challenging to perform stylistic edits and author animations. For example, changing a character's style from relaxed to energetic may inadvertently alter the underlying action, such as turning a walk into running. Therefore, disentangling these aspects is crucial in allow the action to remain fixed while independently changing only its style, enabling precise control for animation tasks like behavioral style transfer.

While several areas in multimodal learning like speech understanding (Mohebbi et al., 2024) and facial animation (Chai et al., 2024) explore disentanglement as part of representation design, motion representations typically model these aspects together and do not isolate style specific traits from content and vice versa. For the task of stylizing content motion given a reference style, methods trained with these representations perform style transfer during generation. Typically, these methods leverage lookup tables to represent fixed styles during training (Aberman et al., 2020; Mughal et al., 2024; Yang et al., 2023) or rely on finetuning low-rank style adapters over pretrained generative models (Sawdayee et al., 2025). As a result, zero-shot generalization of these approaches to unseen styles becomes impossible as stylization requires label input or post-hoc finetuning on labelled data. Another common approach is to perform this transfer during generation through control or guidance

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

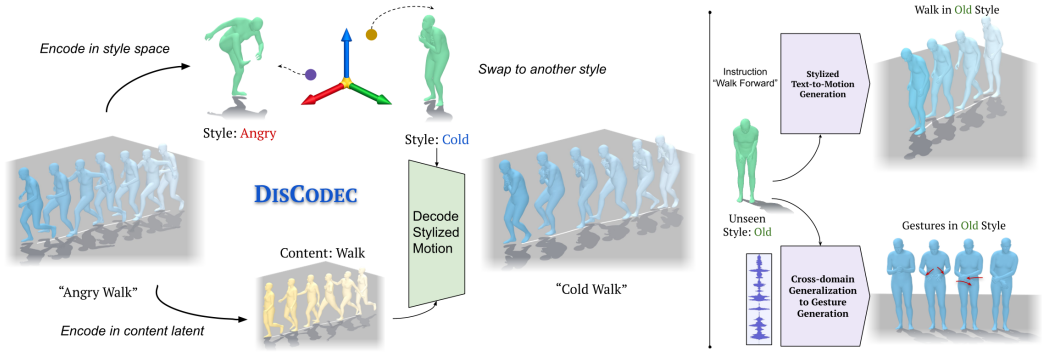


Figure 1: **DISCODEC**: A codec framework to seperate style and content into disentangled latents, which can be used for zero-shot stylization for multimodal generation and cross-task style-transfer.

mechanisms (Zhong et al., 2024; Habibie et al., 2022). However, this approach of achieving style transfer using action-optimized representations binds generation to content motion distribution and hinders style transfer. In contrast, we address the problem at representational level and present a new way to build disentangled motion representations for style and content to achieve high quality zero-shot transfer to reference styles or gestural identities. Having such disentanglement is useful to allow for style or content-specific edits in the latent space of human motions.

To build such a representation space, we propose DISCODEC i.e. a general codec framework to disentangle motion style and content in an abstract latent space. These resulting latents provide separate representations for the two aspects of input motion sequence, that can either be decoded back to stylized outputs or used for motion synthesis methods (Chen et al., 2023) with additional stylization support. Concretely, we aim to perform zero-shot stylized motion generation for two motion domains: text-to-motion synthesis and co-speech gesture synthesis. Given a content input, like instruction or speech, and a reference style, our goal is to synthesize human motions that imbibe the style exhibited in the reference motion, while remaining aligned to the underlying input.

We achieve this by building a codec architecture that encodes motion sequences into an abstract representation and reconstructs it back through a decoder. For this, we introduce a Variational Information Bottleneck (VIB) (Alemi et al., 2017), which removes shared information between the two latents representing different aspects of motion. Furthermore, we make this style latent space more controllable, by taking a parametric approach and factoring the bottlenecked style latents into an orthogonal basis. This lets us explore the style latent space in a procedural manner (by sweeping through the coefficients of the basis set) without even requiring the reference motions. Additionally, the orthonormal style basis acts as a regularization against the memorization of class-specific clusters. Finally, this design also allows one to generate the style latents from a generator, as we show later in section 3.2.

While learning to transfer stylistic patterns is the goal here, we also need to preserve the content adherence instead of blindly copying the style reference. Interestingly, we observe that applying the VIB on content latents does not fully disentangle the content, and causes the style-specific aspects to leak into the content features at low-variance directions. Therefore, we remove these low-variance directions by performing SNR-adaptive Transform Quantization, which rounds off these style-aligned low-SNR directions of the posterior mean. This ensures separation of content from style, directly contributing to better zero-shot transfer to unseen styles. Lastly, we also demonstrate application of our disentangled codec for zero-shot style transfer and style sampling in context of flow-matching driven motion generation paradigm. The proposed disentangled representations automatically provide zero-shot transfer capability through the decoder that simply reconstructs stylized motion, without needing any style adapters or control frameworks during generation. In summary, our contributions are the following:

1. We present **DISCODEC**, a general-purpose codec that disentangles style and content of human motion at a representation level, effectively enabling zero-shot style transfer.
2. We demonstrate the effectiveness of variational information bottleneck in achieving this disentanglement of motion domain and SNR-adaptive Transform Quantization to reduce leakage among the learned content and style latents.
3. We show the zero-shot stylization ability of our codec framework for multimodal motion generation tasks like text-to-motion synthesis and co-speech gesture generation.

2 Related Works

2.1 Disentanglement in Representations

Disentangled representations separate factors of variation, either through unsupervised objectives (Higgins et al., 2017) or through complementary latents in images (Huang et al., 2018) and videos (Denton & Birodkar, 2017). Closer to our setting, speech representations have been split into textual and acoustic features through conditional VIB (Mohebbi et al., 2024). Moreover, facial animation methods disentangle subject-specific speaking style from speech content (Chai et al., 2024; Fu et al., 2024). In comparison, our codec learns both style and content latents jointly from raw human motion, without labels on the style latent or a factor given to the decoder.

However, disentangled latents align with the principal directions in data (Rolinek et al., 2019). Such directions make the posterior mean carry correlations that its samples do not (Bonheme & Grzes, 2023), which we find leak information from style into content. We ground our quantization in transform coding (Goyal, 2001), which quantizes decorrelated coefficients with their own step sizes, and in reverse water-filling that assigns zero rate to components below the distortion level (Cover et al., 1991). Closest to ours, Yang et al. (2020) allocate bits for compression using posterior uncertainty, whereas our framework uses it to set the step of each principal direction and erase the low-SNR directions where style leaks into content. Unlike VQ-VAE (van den Oord et al., 2017) and FSQ (Mentzer et al., 2024), this step follows the measured SNR of each direction.

2.2 Multimodal Behavior Generation

Text-to-motion generation aims to generate a sequence of poses along time by following the underlying instruction (Guo et al., 2022; Tevet et al., 2023; Dabral et al., 2023). Text-driven stylization methods (Zhong et al., 2024; Sawdayee et al., 2025; Raab et al., 2024) are typically built as finetuning-based or inference-time modulation frameworks and therefore, do not address the inherent dichotomy of content and style at motion representation stage. On the contrary, our framework creates disentangled representations of content and styles from raw motion, that can also be used during generation. Co-speech gesture generation methods (Liu et al., 2024; Mughal et al., 2024) typically learn a fixed lookup table of labelled identities that are used at inference for style-specific gesture generation (Yang et al., 2023; Ahuja et al., 2020; Yoon et al., 2020). However, zero-shot personalization remains infeasible on these architectures due to their reliance on predefined identities.

2.3 Stylized Motion Generation

Example Driven Stylization. Reference motions that represent a specific style of motions are used as exemplars to stylize input action-specific motions. Guo et al. (2024) perform this task by learning probabilistic style space and require pretrained latents that are used for disentanglement at post-hoc stage. This can reduce usability at downstream motion generation. In comparison, we approach this idea through a codec that inherently decomposes style and content aspects from raw motion. Kim et al. (2024) utilize Siamese encoders for separately encoding content and styles, followed by Zargarbashi et al. (2026) who apply contrastive loss terms to Residual-VQ levels. Our approach revisits the disentanglement idea with explicit focus on learning a parametric space that can represent unseen styles rather than being aligned to labels.

For speech-to-gesture task, ZeroEGGS Ghorbani et al. (2023) introduces zero-shot emotion stylization on a small emotion dataset. GestureDiffuCLIP (Ao et al., 2023) utilize CLIP embeddings of motion as stylization reference. These approaches summarize style reference into a global vector, entangling its idiosyncratic aspects with transient motion attributes and limiting controllability.

Controllable and Editable Synthesis. Motion generation approaches, both classical (Arikan & Forsyth, 2002; Kovar et al., 2002; Chi et al., 2000) and data-driven Habibie et al. (2022); Ferstl et al. (2021), also provide a way to control their output with low-level motion control like spatial location or motion attributes. SINC (Athanasiou et al., 2023) and FineMoGen Zhang et al. (2023) provide synthesis control via body-part level text prompts. Alexanderson et al. (2020) control output like gesture level, speed, and symmetry in the output. Following this, Yoon et al. (2021) present a toolkit with the control parameters like spatial extent and handedness. Here, control is dependent on prede-

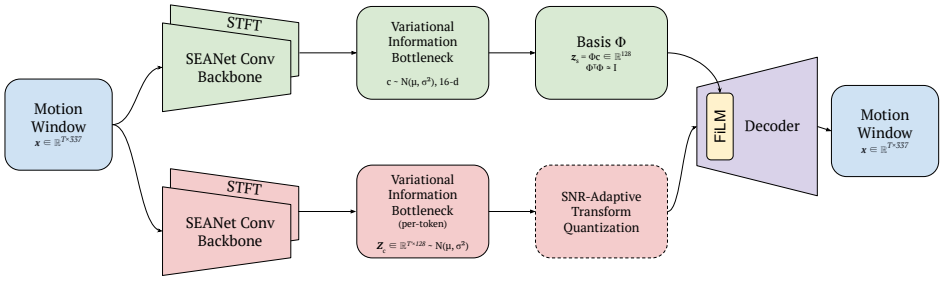


Figure 2: **DISCODEC Architecture**. We encode the input motion into two disentangled latents before decoding them back for reconstruction. The style latents (top branch) are learned through a VIB followed by an orthogonal basis factorization. The content latents (bottom branch) use SNR-adaptive Transform Quantization to prevent the leakage of style into content latents.

fined attributes that let you perform local edits in content. This binds stylization to these attributes and does not learn a representation for idiosyncratic style, disentangled from content details.

Identity-specific Gesture Stylization. Personalized Gesture Generation is a well-studied topic in animation literature (Stone et al., 2004; Noot & Ruttkay, 2004; Neff et al., 2008). More recently, Ahuja et al. (2022) present a low-resource adaption technique where a pretrained generator is fine-tuned on a few minutes of new speaker's motion. This is further extended to a continual learning setting as well (Ahuja et al., 2023). Nevertheless, these approaches focus on learning a specialized model for each unseen identity. Instead, our approach learns to capture style of a new speaker through its representation, which serves as a generalizable proxy to person-specific model.

3 APPROACH

Given full-body human motion as input e.g. *"excited walking"*, our goal is to learn disentangled representations for style and content of this motion. We learn them through an encoder-decoder architecture that utilizes two stream of encoders each belonging to style and content of motion. We apply a Variational Information Bottleneck (VIB) between the two latent representations that isolates style-specific information in the style latent which in this example relates to excitement. In order to allow the content latents to focus on action-specific aspects of motion like walking, we further subject the content latents to SNR-adaptive Transform Quantization (Section 3.1).This provides separate content and style representations, that can be further utilized for generative motion and gesture synthesis (Section 3.2).

Preliminaries: Disentanglement through Variational Information Bottleneck (VIB) Information bottleneck (IB) (Tishby et al., 1999) has been used to disentangle different representations *e.g.* speech features and their content (Mohebbi et al., 2024; Ju et al., 2024) or facial identity and features (Gao et al., 2021). In principle, the goal is to learn a stochastic encoding z that maximizes prediction of target y, while minimizing the information about the input x.

$$\mathcal{L}_{\text{IB}} = \beta I(x, z) - I(z, y) \tag{1}$$

Here I represents mutual information between the two variables and β is the trade-off parameter. In practice, IB loss is intractable and therefore VIB has been proposed (Alemi et al., 2017).

$$\mathcal{L}_{\text{VIB}} = \mathcal{L}_{\text{task}}(y; z) + \beta \, \text{Div}_{\text{KL}}(q(z \mid x) \,\|\, p(z)) \tag{2}$$

The loss $\mathcal{L}_{\text{task}}$ is the conventional reconstruction loss, applied to a neural network decoder to ensure that the latent representation z predicts target y accurately. Further, VIB includes an information loss in the form of KL divergence between the stochastic encoder $q(z \mid x)$ and the marginal $p(z)$, typically assumed to be spherical Gaussian. VIB has been used for one-directional disentangling by (Mohebbi et al., 2024) to learn complementary representation on top of pre-existing representation. In our framework, *we utilize similar VIB to perform bidirectional disentanglement between two representations of style and content* (Section 3.1).

3.1 DISCODEC

The codec takes a motion sequence $\mathbf{x} \in \mathbb{R}^{T \times d}$ as input, containing full-body 6D joint rotations, joint positions, per-frame root velocity and foot contacts. n This input is passed onto two individual encoding streams consisting of style backbone $\mathcal{E}_s(\mathbf{x})$ and content backbone $\mathcal{E}_c(\mathbf{x})$. Each encoding backbone comprises SEANet downsampling backbone (Défossez et al., 2024) and a parallel STFT layer, producing outputs which get passed into a transformer encoder. Both encoding streams utilize a variational information bottleneck (VIB) after their backbone to retain information specific to the corresponding aspect of motion. More specifically, style bottleneck then produces a single vector of basis coefficients $\mathbf{c}_s$ that are transformed into a latent $\mathbf{z}_s \in \mathbb{R}^{128}$, encompassing style for a single motion clip. In parallel, the downsampling content branch applies SNR-Adaptive Transform Quantization on its VIB output to get a per-chunk latent sequence $\mathbf{Z}_c = \{\mathbf{z}_c^1, \mathbf{z}_c^2, \cdots, \mathbf{z}_c^{T\prime}\}, \mathbf{z}_c^i \in \mathbb{R}^{128}$ where $T' = T/M$.

Individual outputs of each encoding stream i.e. $\mathbf{z}_s$ and $\mathbf{Z}_c$, are eventually passed onto the decoder $\mathcal{D}$ that reconstructs stylized motion from $\mathbf{Z}_c$ by using $\mathbf{z}_s$ as conditioning for FiLM layer(Perez et al., 2018). The decoder consists of a transformer and a subsequent upsampling SEANet backbone, similar to encoders. This encoder-decoder architecture is trained through a self-supervised reconstruction objective $\mathcal{L}_{\text{rec}}$ (See Appendix for further details).

Style subspace learning with VIB. Typically, motion codecs are trained using the standard reconstruction objectives (MSE and KL-divergence), as the core goal is to learn a latent representation for motion. However, this objective fails to build a compact latent space that distinguishes between action-specific content and idiosyncratic style in its representation. Therefore, we treat the reconstruction objective as the task loss and introduce a Variational Information Bottleneck on the latents produced by the style backbone $\mathcal{E}_s$, whose stochastic encoder $q(\mathbf{c}_s \mid \mathbf{x}) = \mathcal{N}\left(\boldsymbol{\mu}_s(\mathbf{x}), \operatorname{diag} \boldsymbol{\sigma}_s^2(\mathbf{x})\right)$ produces the style code. The corresponding information loss $\beta_s \mathrm{Div}_{\mathrm{KL}}\left(q(\mathbf{c}_s \mid \mathbf{x}) \,|\, p(\mathbf{c}_s)\right)$ discards information about the input $\mathbf{x}$ beyond what is required for reconstruction.

While creating an information bottleneck can reduce redundant information, the style encoder also needs to learn the structure of stylistic patterns in motion data. A naïve way of incorporating this distinction would be to supervise the style latent space with predefined labels (Aberman et al., 2020), learning class-specific representation structure. In contrast, our goal is to learn a *parametric space* where any new style can be embedded. Therefore, we introduce a learned basis after the VIB, which represents the style latent as a linear combination of 16 learned basis vectors, weighted by the style coefficients $\mathbf{c}_s \in \mathbb{R}^{16}$.

$$\mathbf{z}_s = \Phi \mathbf{c}_s \in \mathbb{R}^{128}, \qquad \text{s.t.} \quad \Phi^\top \Phi \approx \mathbf{I} \tag{3}$$

To further introduce content invariance to style latents, we regularize $\mathbf{z}_s$ with VICReg (Bardes et al., 2021) using two temporal crops of the same clip: the invariance term enforces within-clip consistency, while the variance and covariance terms prevent collapse. This results in a label-free learning paradigm which prevents learning class-specific structure and achieves better zero-shot transfer to unseen styles. Finally, the losses that act on the style stream turn out to be:

$$\mathcal{L}_{\text{style}} = \beta_s \mathrm{Div}_{\mathrm{KL}}(\mathbf{c}_s) + \lambda_s \mathcal{L}_{\text{VICReg}}(\mathbf{z}_s, \mathbf{z'}_s) + \lambda_{\text{o}} \left\| \Phi^\top \Phi - \mathbf{I} \right\|^2 \tag{4}$$

Content space learning with SNR-adaptive Transform Quantization. Action-specific parts of motion like instruction-driven movement or speech-related gestures describe its content. Therefore, a disentangled representation must preserve content in motion while also preventing the learning of the same information as $\mathbf{z}_s$. Content encoding stream downsamples motion to get a latent sequence $\mathbf{Z}_c = \{\mathbf{z}_c^1, \ldots, \mathbf{z}_c^t, \ldots, \mathbf{z}_c^{T'}\}$. Learning an entangled latent through vanilla reconstruction loss without the information bottleneck increases leakage of style information in features, which in turn hurts unseen style transfer (refer to Section 5 for analysis). Therefore, a VIB is applied on each latent token $\mathbf{z}_c^t$ in the content sequence as well, with encoder $q(\mathbf{z}_c \mid \mathbf{x}) = \mathcal{N}\big(\boldsymbol{\mu}_c, \operatorname{diag} \boldsymbol{\sigma}_c^2\big)$, where $\boldsymbol{\mu}_c$ is the posterior mean and $\boldsymbol{\sigma}_c$ the bottleneck noise. The content latent therefore remains continuous with a marginal prior applied per token.

Interestingly with a continuous latent sequence $\mathbf{Z}_c$, style specific information still resides in low-variance directions of the 128-dimensional token, indicating weak disentanglement. This reflects a known property of continuous VAE latents: the posterior aligns with principal directions, ordered by a variance-to-noise criterion below which components collapse to the prior (Rolinek et al., 2019;

Algorithm 1 SNR-adaptive transform quantization of a content token.

Require: content token $\mathbf{z}_c^t = \boldsymbol{\mu}_c + \boldsymbol{\sigma}_c \odot \boldsymbol{\epsilon}$,
step scale δ,
$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
(index t omitted below)

$V \boldsymbol{\Lambda} V^\top \leftarrow \mathrm{Cov}(\boldsymbol{\mu}_c)$ ▷ EVD; $s_j = \sqrt{\lambda_j}$
$n_j \leftarrow \left(\mathbf{v}_j^\top \mathrm{diag}(\bar{\boldsymbol{\sigma}}^2)\, \mathbf{v}_j\right)^{1/2}$ ▷ noise scale along $\mathbf{v}_j$
$\Delta_j \leftarrow \delta\, n_j$ ▷ constant step in SNR units
$\boldsymbol{\eta} \leftarrow \mathbf{z}_c - \boldsymbol{\mu}_c$ ▷ sampling noise; $\mathbf{0}$ at inference
$\mathbf{u} \leftarrow V^\top (\boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}})$ ▷ decorrelate the mean
$\hat{u}_j \leftarrow \Delta_j \lfloor u_j / \Delta_j \rceil$ ▷ $|u_j| < \Delta_j/2 \Rightarrow \hat{u}_j = 0$
$\hat{\boldsymbol{\mu}}_c \leftarrow V \hat{\mathbf{u}} + \bar{\boldsymbol{\mu}}$
$\boldsymbol{\mu}_c \leftarrow \boldsymbol{\mu}_c + \mathrm{sg}\left[\hat{\boldsymbol{\mu}}_c - \boldsymbol{\mu}_c\right]$ ▷ straight-through
return $\mathbf{z}_c \leftarrow \boldsymbol{\mu}_c + \boldsymbol{\eta}$

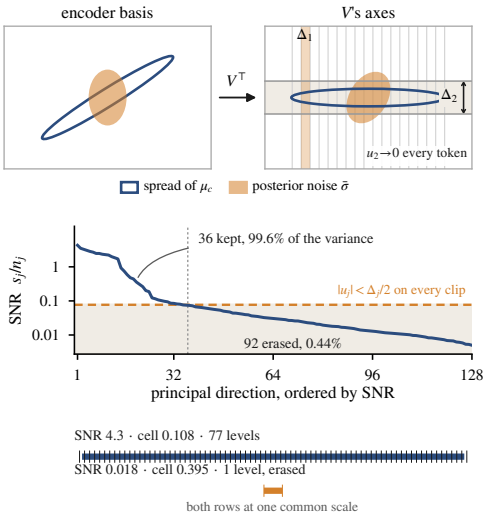


Figure 3: **SNR-adaptive transform quantization. Top (schematic):** V decorrelates the code, and every direction gets its own cell $\Delta_j = \delta n_j$, set by that direction's posterior noise. The code spans many cells along v_1, and along v_2 it lies inside a single cell, so $u_2 \to 0$ for every token, collapsing token to $\bar{\boldsymbol{\mu}}$. **Bottom (measured, $\delta = 0.5$):** 36/128 directions survive carrying 99.6% of the variance.

Lucas et al., 2019). Performing a probing analysis on continuous latents results in straight-forward classification of stylistic patterns present in $\mathbf{Z}_c$ (style←zc: 0.725 in Section 5). Therefore, we mitigate this by introducing SNR-Adaptive Transform Quantization, which erases the low-SNR directions of the posterior mean in which style information resides and achieves better disentanglement of content from style (Fig. 3.1).

VIB already provides insight into how much resolution the channel carries in each dimension of token $\mathbf{z}_c^t$. However, the raw coordinates of $\mathbf{z}_c^t$ are correlated, so quantizing them independently is ill-posed. We therefore quantize in the eigenbasis of $\mathrm{Cov}(\boldsymbol{\mu}_c)$, centered at the running mean $\bar{\boldsymbol{\mu}}$, where the coordinates are uncorrelated (Huang & Schultheiss, 1963; Cover et al., 1991). Along each eigenvector $\mathbf{v}_j$, there exists a signal scale $s_j = \sqrt{\lambda_j}$ i.e. the standard deviation of the posterior mean in that direction, and a noise scale n_j i.e. the standard deviation of the batch-averaged bottleneck noise $\bar{\boldsymbol{\sigma}}$ resolved in the same direction, thus making $\mathrm{SNR}_j = s_j / n_j$. We set the quantization step to each direction's own noise, $\Delta_j = \delta\, n_j$ so directions whose SNR_j falls below a threshold proportional to δ are erased, and at neglible reconstruction cost since distortion below the noise is invisible to the trained decoder(Algorithm 1). Therefore, a direction in eigenspace whose coordinates never exceed $\Delta_j/2$ rounds to zero on every token and collapses to $\bar{\boldsymbol{\mu}}$, while stronger directions are perturbed by at most half a cell. We utilize straight-through estimator sg (Bengio et al., 2013) during encoder training. The eigenbasis V, the center $\bar{\boldsymbol{\mu}}$ and the step Δ_j are built from detached running averages of $\mathrm{Cov}(\boldsymbol{\mu}_c)$, $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c^2$, so the model cannot adjust its own quantization to satisfy the objective. We also apply a cycle term to make the content latent invariant to the style $\mathcal{L}_{\mathrm{cyc}}^c = \left\| \mathcal{E}_c\big(\mathcal{D}(\mathbf{Z}_c, \mathbf{z}_s')\big) - \mathrm{sg}[\mathcal{E}_c(\mathbf{x})] \right\|_2^2$, where $\mathbf{z}_s'$ is a style code the clip did not come from. Refer to Appendix for implementation details.

Stylization at Decoder After the two encoding streams, we end up with two representations: $\mathbf{Z}_c$ for content and $\mathbf{z}_s$ for style of motion. The decoder utilizes $\mathbf{z}_s$ to condition the prediction of reconstructed motion by upsampling $\mathbf{Z}_c$ to produce $\hat{\mathbf{x}}$ This can achieve zero-shot style transfer to unseen styles without any post-hoc guidance or adaptation to the decoding stage. Not limited to this, the resulting representations for style and motion can be further applied to perform motion synthesis as well as style generation through the learned subspace.

3.2 Motion Generation with DisCodec

Flow-matching based Generator. To demonstrate the validity of our learned motion representation as plausible alternative to standard motion latents, we train a standard DiT (Peebles & Xie, 2023) using DisCodec's style and content latents as input. Concretely for a given motion, we concatenate style basis coefficients with sequence of pre-quantization content latents to form the input sequence for DiT, which is trained with conditional flow-matching objective (Lipman et al., 2023). More importantly, a single generator learns the joint distribution of style and content as well as

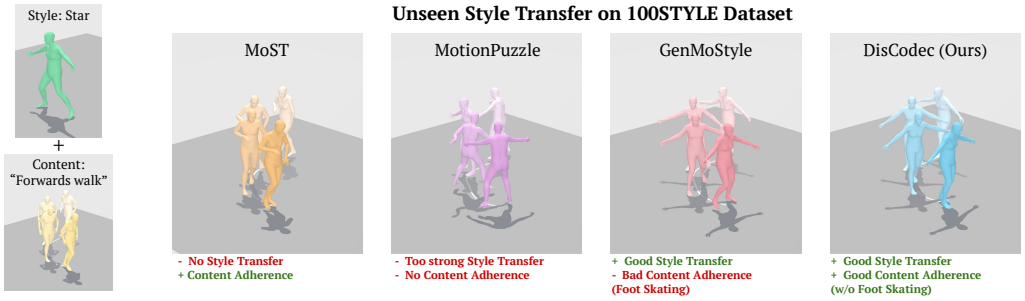


Figure 4: **Qualitative comparison between frames of Reference-based Stylization**

each stream conditioned on the other. For every training sample, the generator either generates both streams or generates one given the other; the given stream is provided as ground truth, flagged with a learned embedding, and excluded from the flow-matching loss. Lastly, we utilize cross-attention for conditioning the task-specific generation. We train two versions: one for text-driven motion generation (T2M) and the other for co-speech gesture generation (S2G).

Stylizing existing T2M Approaches. In addition to this, we also adapt an existing text-to-motion generation framework ActionPlan (Nazarenus et al., 2026) to demonstrate the generalization capability of the proposed codec to existing T2M frameworks. ActionPlan is also a diffusion framework that combines real-time streaming with offline motion generation from text. We utilize only content stream of encoded motions as input data to retrain ActionPlan, keeping rest of method unchanged. Note that, since our corpora lack frame-level text latents, we use its text-latent-free regime, and style is applied by our decoder from a reference or sampled style code at the decoding stage after generation. Refer to Appendix for details on generators.

4 Experiments

To evaluate the style transfer capabilities of our approach, we demonstrate its performance across three axes. Firstly, we compare our method on the *reference-based motion stylization task*, where the goal is to stylize input motion according to a reference clip, while keeping its content intact. Secondly, we apply DisCodec's encoding capabilities to stylized *text-to-motion generation task* by evaluating our synthesis models against generation-based stylization baselines. Here, we perform stylization at the codec level by fixing the reference style latent during generation and generating text-driven motion. Lastly, we test our method on co-speech generation and demonstrate the capability of our model to transfer *identity-specific idiosyncratic behavior* for co-speech gesture generation models, where we treat personal identity as style. For this task, our goal is to perform transfer of speaker-identifying traits to the generated gestures by extracting reference speaker's style through DisCodec and stylizing gesture generation using that.

Datasets. For reference-based motion stylization, we benchmark our approach on 100STYLE (Mason et al., 2022). For text-to-motion stylization, we utilize HumanML3D (Guo et al., 2022) along with 100STYLE and PerMo (Kim et al., 2025), forming a union of three datasets. For the seen classes, train and test are split at the level of recordings, so no recording contributes to both. For gestural identity transfer, we use BEATX (Liu et al., 2024) and SeamlessInteraction (Agrawal et al., 2025). Refer to appendix for details.

Baselines. For Motion2Motion Stylization baselines, we compare our method with MoST(Kim et al., 2024), MotionPuzzle(Jang et al., 2022), GenmoStyle(Guo et al., 2024) by retraining each baseline for 100STYLE and PerMo datasets. To compare DisCodec's application to generation-based text-to-motion baselines, we benchmark our synthesis models against SMooDi (Zhong et al., 2024), LoRA-MDM (Sawdayee et al., 2025), MoMo (Raab et al., 2024). For speaker-identity transfer task, we compare our zero-shot stylization approach with lookup-based gesture generation methods like EMAGE Liu et al. (2024) and RAG-Gesture Mughal et al. (2025).

4.1 Reference-based Stylization

Table 1 shows the comparison of DisCodec with state-of-the-art baselines trained on 100STYLE. We observe that our method retains reconstruction fidelity compared to the baselines, shown by

Table 1: **Motion-to-motion style transfer on 100STYLE.** "style ← $\mathbf{z}_c$" and "content ← $\mathbf{z}_s$" represent linear probe accuracy to measure forward and reverse leakage respectively.

Method	Seen SRA ↑	CRA ↑	Unseen SRA ↑	FID ↓	FSR ↓	style ← $\mathbf{z}_c$ ↓	content ← $\mathbf{z}_s$ ↓
GT	0.967	0.601	1.000	0.51	0.216	–	–
MoST	0.017	0.465	0.000	1.80	0.298	0.882	0.699
GenMoStyle	0.734	0.257	**0.447**	1.19	0.524	**0.052**	0.701
Motion Puzzle	0.413	0.142	0.300	3.55	0.650	0.685	0.709
DISCODEC (union)‡	0.547	0.342	0.279	1.15	**0.186**	0.202	**0.516**
DISCODEC (100STYLE)	**0.756**	**0.467**	0.161	**1.05**	0.246	0.138	**0.516**

Table 2: **Stylized text-to-motion under SMooDi's protocol.** SRA: Style Recognition Accuracy, LoRA-MDM and SMooDi are trained on every style, so their SRA is not zero-shot.

Method	Zero-shot	SRA ↑	Unseen SRA ↑	R-Prec@1 ↑	MM-Dist ↓	FID ↓	Diversity →	Skate ↓
GT	–	0.996	0.989	0.486	3.00	0.00	9.63	0.077
LoRA-MDM	✗	0.192	0.069	**0.507**	**3.09**	**0.53**	9.40	**0.064**
SMooDi	✗	0.182	0.089	0.366	3.98	1.30	**9.47**	0.140
MoMo	✓	**0.597**	**0.523**	0.101	7.74	23.89	6.01	0.477
Ours, DiT	✓	0.347	0.267	0.270	4.94	7.65	7.38	0.068
Ours, ActionPlan	✓	0.343	0.235	0.256	4.99	6.80	7.43	0.072

Table 3: **Stylized text-to-motion on PerMo.** PRA :Persona recognition accuracy, Personabooth trains with all styles.

Method	Unseen ref.	R-Prec@1 ↑	MM-Dist ↓	FID ↓	Diversity →	Skate ↓	PRA (%) ↑
GT	–	0.486	3.00	0.00	9.63	0.077	–
PersonaBooth	✗	0.077	7.48	4.07	7.49	0.074	**28.32**
Ours, DiT	✓	**0.342**	**4.21**	**2.46**	**7.83**	**0.036**	8.98
Ours, ActionPlan	✓	0.316	4.42	2.81	7.68	**0.036**	9.18

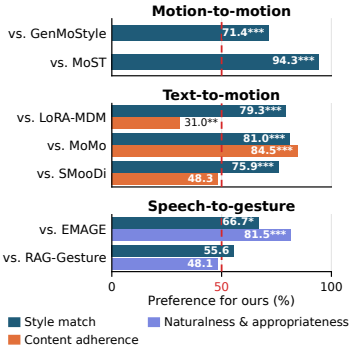


Figure 5: **Perceptual Evaluation.** (*p<.05, **:p<.01, ***p<.001).

the lowest FID and foot-skate ratio (FSR). Interestingly, MoST fails to transfer style completely by copying the content, while MotionPuzzle and GenMoStyle (unsupervised) copy the style reference's motion and struggle to keep content-specific actions, as shown by low CRA and high FID. Our method balances content adherence and stylization, achieving high SRA along with the best FID and CRA. More importantly, we perform probing analysis on the content and style latents of each method with linear classifiers to measure disentanglement. GenMoStyle only separates style from content latents, whereas our approach also reduces the reverse leakage of content into $\mathbf{z}_s$, achieving the most balanced bi-directional disentanglement. Figure 4 further illustrates these patterns. Importantly, we perform a perceptual evaluation (Fig. 5) to compare stylization capability and find that our approach outperforms baselines by wide margin. For consistency with the baselines, metrics in this domain exclude root prediction. Lastly, we train two versions of DISCODEC: one on 100STYLE alone and one on the union of all text-to-motion stylization datasets for downstream use.

4.2 TEXT-TO-MOTION STYLIZATION

To further evaluate our representation-forward stylization approach, we compare the generation frameworks trained with DISCODEC's latents against the text-to-motion stylization baselines trained on 100STYLE dataset (Table 2). The baselines range from methods with high stylization performance but worst content adherence (MoMo) to methods that fail to generalize to unseen styles but perform the best on FID (SMooDi, LoraMDM). DISCODEC provides the most balanced performance amongst all showing high zero-shot stylization performance while also keeping content adherence. Our zero-shot generators lie in the middle of the pareto front: far higher unseen-style SRA than LoRA-MDM and SMooDi at a fraction of MoMo's FID (see Supplementary Material). We follow SMooDi's protocol for evaluation and use their released evaluation models for SRA and FID. Note that content adherence metrics like FID are calculated against HumanML3D data distribution, which is not the same distribution as the target style distribution of 100STYLE. Therefore, it is natural for FID to increase as the style performance improves. Furthermore, we also compare with PersonaBooth Kim et al. (2025) on PerMo dataset's protocol. PersonaBooth trains with all personas in their dataset and, therefore, achieves the best PRA. In comparison, our zero-shot approach achieves reasonable PRA, while outperforming the baseline across all content adherence metrics. Since the quantitative metrics can be flawed in terms of holistic evaluation, we conduct a perceptual evaluation on the stylization and content adherence capability on unseen styles (Fig. 5). We find that

our approach outperforms all finetuning-based or zero-shot baselines for style transfer. Note, that our method outperforms MoMo on both style transfer and content adherence. In terms of content adherence, finetuning-based LoRA-MDM understandably performs better.

4.3 Speech-to-Gesture Personalization

We evaluate our approach on speaker-specific identity transfer against synthesis baselines that produce person-specific motions by conditioning on identity labels from learned lookup embeddings.

We observe better content and identity adherence through a zero-shot approach, performing on par with methods that trained on full identity set. Note that, we report Identity Recognition Accuracy (IRA) on all seen identity to be consistent with evaluation of look-up based approaches. RAG-gesture performs better at FID because it inserts semantic retrievals into output from GT distribution. However, to properly evaluate generalization to unseen identities, we perform a perceptual evaluation where references from completely unseen identities are shown and evaluators are asked to identify the person from their gesture profile. We observe our zero-shot method performs on-par with baselines that are trained on held-out identities, importantly outperforming them for identity preservation (Fig. 5)

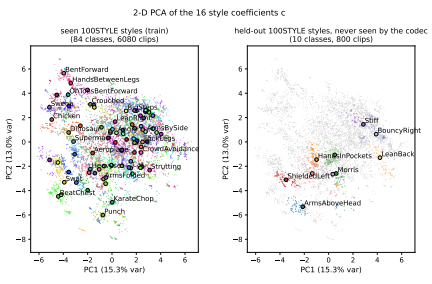
Method	FGD ↓	BC ↑	L1div →	IRA ↑
GT	-	0.424	7.24	1.000
EMAGE	1.80	0.318	7.52	**1.000**
RAG-Gesture (LLM-based)	**1.31**	0.510	10.80	0.898
Ours, DiT	1.71	**0.387**	**7.26**	**1.000**

5 Ablations and Analysis

Codec	MPJPE (mm) ↓	Seen SRA ↑	Unseen SRA ↑	CRA ↑	style ← z_c ↓	content ← z_s ↓
DisCodec	33.5	0.479	0.184	0.374	0.193	0.521
w/o SNR-Q (cont. z_c)	43.9	0.422	0.233	0.314	0.725	0.389
w/o SNR-Q (FSQ on z_c)	35.9	0.384	0.202	0.258	0.185	0.575
w/o VICReg	47.5	0.061	0.015	0.344	0.236	0.306
w/o cycle loss	45.6	0.390	0.141	0.376	0.110	0.483
w/o style VIB	43.9	0.457	0.201	0.311	0.095	0.605
w/o content VIB†	37.2	0.161	0.060	0.300	0.844	0.272

Ablations of Design Choices. While ablating the proposed architectural components, we observe that replacing SNR Adaptive Transform Quantization with continuous latents clearly results in high leakage of style information in $\mathbf{z}_c$. Moreover, replacing it with another rounding quantizer like FSQ (Mentzer et al., 2024) majorly reduces content information in outputs. Importantly, content-stream VIB keeps style out of $\mathbf{z}_c$, while style VIB guards the reverse direction. Lastly, we observe that VICReg is quite important for r style transfer and disentanglement. Full method also achieves the best MPJPE.

Analysis of Style Subspace. We analyze the learned style subspace by plotting PCA of 16 style coefficients Section 5. Interestingly, the subspace exhibits meaningful relations between style classes without collapsing to label specific clusters. We observe that styles like "KarateChop" and "Punch" lie close together as well as styles that are bent forward e.g. "HandsBetweenLegs" and "BentForward". Importantly, unseen styles, when projected into this subspace, also exhibit similar semantic structure: "SheildedLeft" lies close to "Swat" which contains pose with one arm bent and the other straight.



6 Conclusion

In this work, we showed the effectiveness of variational information bottleneck in isolating the style of full-body human motion from the base motion. More importantly, we introduced a parametric basis of the style space by decomposing the style embedding into a learnable basis set. As our evaluations show, this design meaningfully disentangles style from the rest of the motion, enabling analyses and applications in style transfer. We also presented an SNR-adaptive transform quantization approach for encoding motion content. We hope to demonstrate further downstream applications of our disentangled motion representations in the future.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, assist with translation,support qualitative and thematic data analysis. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, interpret results. and clean and reformat dataset is not applicable to this work. Additionally, we used generative AI tools to create or edit software code, creation of artifacts. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. ACM Transactions on Graphics (TOG), 39(4):64, 2020.

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. arXiv preprint arXiv:2506.22554, 2025.

Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. 2020.

Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. Low-resource adaptation for personalized co-speech gesture generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20566–20576, 2022. doi: 10.1109/CVPR52688.2022.01991.

Chaitanya Ahuja, Pratik Joshi, Ryo Ishii, and Louis-Philippe Morency. Continual learning for personalized co-speech gesture generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 20893–20903, 2023.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In 5th International Conference on Learning Representations (ICLR), 2017.

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. Computer Graphics Forum, 39(2):487–496, 2020. doi: 10.1111/cgf.13946. Eurographics 2020.

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

Okan Arikan and D. A. Forsyth. Interactive motion generation from examples. In Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), pp. 483–490. ACM, 2002. doi: 10.1145/566570.566606.

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9950–9961. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00916.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906, 2021.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432, 2013.

Lisa Bonheme and Marek Grzes. Be more active! Understanding the differences between mean and sampled representations of variational autoencoders. Journal of Machine Learning Research, 24 (324):1–30, 2023.

Yujin Chai, Tianjia Shao, Yanlin Weng, and Kun Zhou. Personalized audio-driven 3d facial animation via style-content disentanglement. IEEE Transactions on Visualization and Computer Graphics, 30, 2024. doi: 10.1109/TVCG.2022.3230541.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18000–18010, 2023.

Diane Chi, Monica Costa, Liwei Zhao, and Norman Badler. The EMOTE model for effort and shape. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), pp. 173–182. ACM Press/Addison-Wesley, 2000. doi: 10.1145/344779.352172.

Thomas M Cover, Joy A Thomas, and John Kieffer. Elements of information theory, volume 2. wiley New York, 1991.

Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. 2023.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. arXiv preprint arXiv:2410.00037, 2024.

Emily L. Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. Expressgesture: Expressive gesture generation from speech through database matching. Computer Animation and Virtual Worlds, 32(3-4):e2016, 2021.

Hui Fu, Zeqing Wang, Ke Gong, Keze Wang, Tianshui Chen, Haojie Li, Haifeng Zeng, and Wenxiong Kang. Mimic: Speaking style disentanglement for speech-driven 3D facial animation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 1770–1777, 2024. doi: 10.1609/aaai.v38i2.27945.

Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3404–3413, 2021. doi: 10.1109/CVPR46437.2021.00341.

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. Computer Graphics Forum, 42 (1):206–216, 2023.

Vivek K. Goyal. Theoretical foundations of transform coding. IEEE Signal Processing Magazine, 18(5):9–21, 2001. doi: 10.1109/79.952802.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. 2022.

Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. Generative human motion stylization in latent space. In The Twelfth International Conference on Learning Representations (ICLR), 2024.

Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In ACM SIGGRAPH 2022 conference proceedings, 2022.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations (ICLR), 2017.

J Huang and P Schultheiss. Block quantization of correlated gaussian random variables. IEEE Transactions on Communications Systems, 11(3):289–296, 1963.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. ACM Transactions on Graphics (TOG), 41(3):1–16, 2022.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In International Conference on Machine Learning (ICML), 2024.

Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. Most: Motion style transformer between diverse action contents. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1705–1714. IEEE, 2024.

Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, et al. Personabooth: Personalized text-to-motion generation. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22756–22765. IEEE, 2025.

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), pp. 473–482. ACM, 2002. doi: 10.1145/566570.566605.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. 2023.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. 2024.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Don't blame the ELBO! A linear VAE perspective on posterior collapse. In Advances in Neural Information Processing Systems (NeurIPS), volume 32, 2019.

Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 5(1):1–18, 2022.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In International Conference on Learning Representations (ICLR), 2024.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, Afra Alishahi, and Ivan Titov. Disentangling textual and acoustic features of neural speech representations. arXiv preprint arXiv:2410.03037, 2024.

M. Hamza Mughal, Rishabh Dabral, Merel C. J. Scholman, Vera Demberg, and Christian Theobalt. Retrieving semantics from the deep: an rag solution for gesture synthesis. In Computer Vision and Pattern Recognition (CVPR), 2025.

Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. 2024.

Eric Nazarenus, Chuqiao Li, Yannan He, Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Actionplan: Future-aware streaming motion synthesis via frame-level action planning. arXiv preprint arXiv:2603.13500, 2026.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics, 27(1): 1–24, 2008.

Han Noot and Zsófia Ruttkay. Gesture in style. In Antonio Camurri and Gualtiero Volpe (eds.), Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop (GW 2003), volume 2915 of Lecture Notes in Computer Science (LNAI), pp. 324–337. Springer, Berlin, Heidelberg, 2004. doi: 10.1007/978-3-540-24598-8_30.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4195–4205, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

Sigal Raab, Inbar Gat, Nathan Sala, Guy Tevet, Rotem Shalev-Arkushin, Ohad Fried, Amit H. Bermano, and Daniel Cohen-Or. Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer. In SIGGRAPH Asia 2024 Conference Papers, 2024. doi: 10.1145/3680528.3687579.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue PCA directions (by accident). In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12406–12415, 2019.

Haim Sawdayee, Chuan Guo, Guy Tevet, Bing Zhou, Jian Wang, and Amit H. Bermano. Dance like a chicken: Low-rank stylization for human motion diffusion. arXiv preprint arXiv:2503.19557, 2025.

Matthew Stone, Douglas DeCarlo, Insuk Oh, Christian Rodriguez, Adrian Stere, Alyssa Lees, and Christoph Bregler. Speaking with hands: Creating animated conversational characters from recordings of human performance. ACM Transactions on Graphics, 23(3):506–513, 2004. doi: 10.1145/1015706.1015753. SIGGRAPH 2004.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In The Eleventh International Conference on Learning Representations (ICLR), 2023.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, pp. 368–377, 1999.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017.

Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI), pp. 5860–5868, 2023. doi: 10.24963/ijcai.2023/650.

Yibo Yang, Robert Bamler, and Stephan Mandt. Variational Bayesian quantization. In Proceedings of the 37th International Conference on Machine Learning (ICML), volume 119 of PMLR, pp. 10670–10680, 2020.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics, pp. 1–16, 2020.

Youngwoo Yoon, Keunwoo Park, Minsu Jang, Jaehong Kim, and Geehyuk Lee. SGToolkit: An interactive gesture authoring toolkit for embodied conversational agents. In Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST), pp. 826–840. ACM, 2021. doi: 10.1145/3472749.3474789.

Fatemeh Zargarbashi, Dhruv Agrawal, Jakob Buhmann, Martin Guay, Stelian Coros, and Robert W Sumner. Vq-style: Disentangling style and content in motion with residual quantized representations. In Computer Graphics Forum, pp. e70377. Wiley Online Library, 2026.

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. FineMoGen: Fine-grained spatio-temporal motion generation and editing. In Advances in Neural Information Processing Systems 36 (NeurIPS), 2023.

Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. SMooDi: Stylized motion diffusion model. In Proceedings of the European Conference on Computer Vision (ECCV), 2024. doi: 10.1007/978-3-031-73232-4_23.