Hierarchical Enhancement of Semantic Priors for Disentangled Text-Driven Motion Generation

Supplementary Material

1. Appendix A

Differentiation from Standard AG-VAE: Although Gaussian Mixture Variational Autoencoders have been previously investigated in various domains [1], our proposed Adaptive Gaussian VAE (AG-VAE), unlike standard VAEs that treat all dimensions equally, each mixture component in AG-VAE is explicitly conditioned on the skeletal topology through STConv encoding. This ensures that the learned latent representations respect biomechanical constraints and joint dependencies. We depart from static mixture priors by introducing temporally adaptive mixture weights $\pi_k(t)$ that evolve based on motion characteristics. This dynamic formulation captures the non-stationary nature of human movements, where semantic dominance shifts throughout a motion sequence. Rather than assigning a single semantic indicator k to the entire sequence, our hierarchical assignment operates across multiple temporal granularities (joint-level $\rightarrow$ frame-level $\rightarrow$ sequence-level). This multi-scale approach enables fine-grained semantic control and better alignment with the hierarchical structure of both human motion and natural language descriptions. These architectural innovations distinguish AG-VAE from conventional AG-VAE formulations and are specifically designed to address the unique challenges of structured human motion generation.

1.1. Appendix A.1 Theoretical View: Latent Semantic Decomposition (LSD)

Complete ELBO Derivation for VAE-GMM (AG-VAE) The proposed VAE-GMM framework, motivated by the Latent Semantic Decomposition (LSD) view, assumes a prior distribution $p(\mathbf{z}, k)$ over the latent variable $\mathbf{z}$ and the semantic cluster index k, defined as a Gaussian Mixture Model (GMM):

$$p(\mathbf{z}, k) = p(\mathbf{z}|k)p(k) \tag{1}$$

where $p(\mathbf{z}|k)$ is the k-th Gaussian component and $p(k)$ is the categorical prior over clusters.

We aim to maximize the Evidence Lower Bound ($\mathcal{L}$) on the marginal log-likelihood $\log p(\mathbf{x})$.

1. Definition of the Lower Bound

Starting from the marginal log-likelihood:

$$\log p(\mathbf{x}) = \log \sum_{k=1}^{K} \int p(\mathbf{x}, \mathbf{z}, k) d\mathbf{z} \tag{2}$$

We introduce the variational distribution $q_\phi(\mathbf{z}, k|\mathbf{x}) = q_\phi(\mathbf{z}|\mathbf{x}, k)q_\phi(k|\mathbf{x})$ and apply Jensen's inequality:

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z},k|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}, k)}{q_\phi(\mathbf{z}, k|\mathbf{x})} \right] \triangleq \mathcal{L} \tag{3}$$

2. Expansion of the ELBO

The ELBO ($\mathcal{L}$) can be expanded by decomposing the joint probability $p(\mathbf{x}, \mathbf{z}, k) = p(\mathbf{x}|\mathbf{z}, k)p(\mathbf{z}, k)$ and the variational distribution $q_\phi(\mathbf{z}, k|\mathbf{x})$:

$$\begin{aligned}
\mathcal{L} =& \mathbb{E}_{q_\phi(\mathbf{z},k|\mathbf{x})} [\log p(\mathbf{x}|\mathbf{z}, k) \\
&+ \log p(\mathbf{z}, k) - \log q_\phi(\mathbf{z}, k|\mathbf{x}) \\
=& \mathbb{E}_{q_\phi(\mathbf{z},k|\mathbf{x})} [\log p(\mathbf{x}|\mathbf{z}, k)] \\
&- \mathbb{E}_{q_\phi(\mathbf{z},k|\mathbf{x})} \left[\log \frac{q_\phi(\mathbf{z}, k|\mathbf{x})}{p(\mathbf{z}, k)} \right] \\
=& \mathbb{E}_{q_\phi(\mathbf{z},k|\mathbf{x})} [\log p(\mathbf{x}|\mathbf{z}, k)] \\
&- D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}, k|\mathbf{x}) \| p(\mathbf{z}, k) \right) \quad \textbf{(Step A)}
\end{aligned}$$

3. Decomposition of the KL Term

The core of the VAE-GMM formulation is the decomposition of the joint KL divergence into a KL term for the cluster assignments (k) and a conditional KL term for the latent variable ($\mathbf{z}$):

Using the chain rule for probability:

$$\frac{q_\phi(\mathbf{z}, k|\mathbf{x})}{p(\mathbf{z}, k)} = \frac{q_\phi(\mathbf{z}|\mathbf{x}, k)q_\phi(k|\mathbf{x})}{p(\mathbf{z}|k)p(k)} = \frac{q_\phi(\mathbf{z}|\mathbf{x}, k)}{p(\mathbf{z}|k)} \cdot \frac{q_\phi(k|\mathbf{x})}{p(k)}$$

Substituting this back into the expectation in Step A:

$$\begin{aligned}
& D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}, k|\mathbf{x}) \| p(\mathbf{z}, k) \right) \\
&= \mathbb{E}_{q_\phi(k|\mathbf{x})} \left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},k)} \left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x}, k)}{p(\mathbf{z}|k)} \right] \right] \\
&\quad + \mathbb{E}_{q_\phi(k|\mathbf{x})} \left[\log \frac{q_\phi(k|\mathbf{x})}{p(k)} \right] \\
&= \mathbb{E}_{q_\phi(k|\mathbf{x})} \left[D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}, k) \| p(\mathbf{z}|k) \right) \right] \\
&\quad + D_{\mathrm{KL}} \left(q_\phi(k|\mathbf{x}) \| p(k) \right) \quad \textbf{(Step B)}
\end{aligned}$$

4. Final ELBO Expression

Substituting the decomposed KL term (Step B) back into the ELBO (Step A), we obtain the final expression for the VAE-GMM ELBO, which serves as the objective function for training the AG-VAE:

$$\begin{aligned}
\mathcal{L} =& \mathbb{E}_{q_\phi(\mathbf{z},k|\mathbf{x})} [\log p(\mathbf{x}|\mathbf{z}, k)] \\
&- D_{\mathrm{KL}} \left(q_\phi(k|\mathbf{x}) \| p(k) \right) \\
&- \mathbb{E}_{q_\phi(k|\mathbf{x})} \left[D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}, k) \| p(\mathbf{z}|k) \right) \right]
\end{aligned}$$

1.2. Appendix A.2 Theoretical Comparison with GMM and GMVAE

The Adaptive Gaussian VAE (AG-VAE) builds upon the Gaussian Mixture VAE (GMVAE) foundation but introduces critical structure-awareness and temporal adaptability necessary for complex human motion modeling. This section theoretically contrasts AG-VAE with its predecessors.

A.2.1 Comparison with Standard Gaussian Mixture Models (GMM) Standard GMMs are fundamentally generative density estimators defined as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{4}$$

where parameters are typically optimized via Expectation-Maximization (EM).

A.2.2 Comparison with Gaussian Mixture VAE (GMVAE) GMVAE is a direct VAE extension that uses a static GMM prior. The primary distinction of AG-VAE lies in conditioning the prior on the motion structure and making the cluster assignments time-adaptive.

1. Prior Conditioning: Structure-Awareness

GMVAE assumes a generic, data-independent prior for all components:

$$p(\mathbf{z}, k)_{\text{GMVAE}} = p(k)\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \tag{5}$$

The component means and covariances are learned implicitly via the encoder's output, but the prior components $p(\mathbf{z}|k)$ themselves are simple isotropic Gaussians centered at the origin.

In contrast, AG-VAE defines the component parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k)$ as functions of the skeletal topology (S), learned via the network structure (e.g., Graph Convolutions), ensuring that the disentangled latent subspaces respect anatomical dependencies:

$$p(\mathbf{z}, k)_{\text{AG-VAE}} = \pi_k(t, S)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k(S), \boldsymbol{\Sigma}_k(S, T)) \tag{6}$$

This structure-aware prior fundamentally improves the semantic separability of the latent space.

2. Temporal Adaptivity: Dynamic Semantic Switching

The variational posterior inference in GMVAE is typically static (sequence-level), meaning the cluster assignment k is inferred once for the entire sequence $\mathbf{x}$:

$$q_\phi(k|\mathbf{x})_{\text{GMVAE}} \approx \text{Categorical}(\gamma_1, \ldots, \gamma_K) \tag{7}$$

This forces a complex, long motion (e.g., walking $\rightarrow$ turning $\rightarrow$ stopping) to be compressed into a single, often ambiguous, semantic mode k.

AG-VAE, however, leverages its skeleton-temporal convolutions to infer time-adaptive soft assignments $q_\phi(k|\mathbf{x}_t)$ for each time step t, as well as time-adaptive mixture weights $\pi_k(t, S)$ in the prior:

$$q_\phi(k|\mathbf{x})_{\text{AG-VAE}} \approx \prod_{t=1}^{N} \text{Categorical}(\gamma_{1,t}, \ldots, \gamma_{K,t}) \tag{8}$$

This allows the model to perform multi-granularity semantic switching along the sequence, mapping transitions and short gestures to distinct, dynamically active clusters, which is essential for learning a truly disentangled latent semantic decomposition (LSD) of motion.

From Standard VAE to Mixture-based Representation In a conventional VAE, the latent prior is modeled as an isotropic Gaussian:

$$p(z) = \mathcal{N}(z \mid 0, I), \tag{9}$$

and the encoder learns a posterior distribution $q_\phi(z \mid x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$. However, such unimodal prior is insufficient for human motion, where multimodal transitions and discrete action intentions coexist in the continuous space. Empirically, this limitation leads to posterior collapse and semantic entanglement, where temporally distant but semantically dissimilar motions are mapped into overlapping regions of the latent space.

To alleviate this, we redefine the latent prior as a Gaussian Mixture Model (GMM):

$$p(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z \mid \mu_k, \Sigma_k), \tag{10}$$

where K denotes the number of latent semantic components. This formulation allows each latent component to capture a distinct mode of motion dynamics, and the learned mixture weights π_k naturally reflect the semantic frequency of corresponding motion types in the dataset.

Variational Decomposition of Motion Semantics Given motion x, the encoder produces a posterior $q_\phi(z, c \mid x)$ over latent code z and semantic assignment c:

$$q_\phi(z, c \mid x) = q_\phi(c \mid x), q_\phi(z \mid x, c), \tag{11}$$

where $c \in 1, \ldots, K$ is the latent semantic indicator. Substituting this into the Evidence Lower Bound (ELBO), we obtain:

$$\mathcal{L} = \mathbb{E}_{q_\phi(z,c|x)}[\log p_\theta(x \mid z, c)] - D_{\text{KL}}\left(q_\phi(z, c \mid x) \| p(z, c)\right). \tag{12}$$

This formulation leads to a semantic factorization of the latent space, where $q_\phi(c \mid x)$ captures global motion semantics, and $q_\phi(z \mid x, c)$ models fine-grained variations within each semantic mode.

Such factorization directly supports semantic disentanglement — for example, latent dimensions associated with "walking" and "waving" evolve under different Gaussian components, avoiding interference during reconstruction or generation.

Skeleton-aware Semantic Conditioning Unlike classical GMM-VAEs, we further condition each latent component on skeleton topology and temporal evolution, formalized as:

$$p(z \mid c) = \mathcal{N}(z \mid \mu_c(S), \Sigma_c(T)), \tag{13}$$

where S denotes the skeletal graph and T the temporal index. This skeleton-temporal conditioning ensures that latent semantics are structurally grounded—motions sharing similar joint configurations and temporal rhythms are embedded within the same mixture component. Consequently, our model not only clusters by motion category but also respects structural continuity and biomechanical coherence, improving realism and diversity simultaneously.

Theoretical Insights From a probabilistic perspective, the VAE-GMM can be viewed as performing latent semantic decomposition:

$$z = \sum_{k=1}^{K} \gamma_k(x) z_k, \quad \gamma_k(x) = q_\phi(c = k \mid x), \tag{14}$$

where $z_k \sim \mathcal{N}(\mu_k, \Sigma_k)$. This linear mixture interpretation reveals that each motion is generated as a semantic blend of multiple latent prototypes, offering an interpretable bridge between continuous motion embedding and discrete action understanding.

Moreover, under mild assumptions of conditional independence between joints, it can be shown that the mixture prior minimizes the upper bound of the reconstruction error:

$$\mathbb{E}_{p(x)}[|x - \hat{x}|^2] \leq \sum_{k=1}^{K} \pi_k \mathrm{Tr}(\Sigma_k) + \mathcal{O}(\epsilon). \tag{15}$$

where ϵ denotes higher-order residual terms. Hence, the GMM prior not only enhances semantic expressivity but also tightens the reconstruction bound, leading to empirically observed gains in FID, R-Precision, and diversity.

Discussion: Latent Semantic Decomposition and Structured Motion Manifolds The proposed Latent Semantic Decomposition (LSD) reframes the problem of high-dimensional motion generation from learning a flat, single-mode representation to discovering a structured mixture manifold. This paradigm shift provides substantial theoretical and empirical advantages for modeling human motion dynamics:

1. Enhanced Semantic Interpretability and Disentanglement The hierarchical structure of the VAE-GMM $(\mathbf{z}, k)$ naturally factorizes the motion space. The discrete semantic indicator k (cluster assignment) directly corresponds to distinct motion intentions or structural primitives (e.g., locomotion, gesture, transition). The continuous latent variable $\mathbf{z}$ then models the fine-grained style and variation within that specific semantic mode k. This inherent factorization achieves superior semantic disentanglement, providing a clear mapping between latent variables and physical movement characteristics.

2. Improved Diversity with Controllability By defining the prior as a mixture of Gaussians $p(\mathbf{z}, k)$, the model explicitly captures the multimodal distribution inherent in real-world motion data. This prevents posterior collapse, a common failure mode in standard VAEs—and guarantees that the generative process can sample from multiple distinct regions of the manifold. Furthermore, generation is inherently controllable: Mode Conditioning: By explicitly sampling the cluster k, generation can be conditioned on a specific semantic intent (e.g., forcing the output to be "waving"). Mode Interpolation: Interpolation between different $\boldsymbol{\mu}_k$ allows for the controlled synthesis of natural transitions and compositional movements.

3. Superior Generalization and Bio-mechanical Coherence The structure-aware and time-adaptive nature of our prior ($p(z \mid k, S, T)$) ensures that the latent components respect skeletal topology (S) and temporal rhythms (T). This conditioning embeds bio-mechanical coherence directly into the manifold structure. By learning distinct semantic sub-manifolds, the model exhibits better generalization to unseen compositional patterns, as it can stitch together learned primitives (the components k) rather than relying solely on continuous interpolation within a single, amorphous space.

Through the lens of LSD, our AG-VAE is not merely a technical extension of the standard VAE. It represents a principled reformulation of the latent representation learning process for human motion—one that fundamentally bridges continuous trajectory generation and discrete semantic understanding by modeling motion as a time-varying sequence of semantic components.

1.3. Appendix A.3 Theoretical Justification for the Adaptive Prior

Motivation. While Appendix A.1 and A.2 describe the ELBO formulation and structural extensions of AG-VAE, this section provides a theoretical justification for the necessity and validity of the adaptive prior $p_{\text{adapt}}(\mathbf{z}, k|S, T) = p_{\text{adapt}}(k|S, T)\, p_{\text{adapt}}(\mathbf{z}|k, S, T)$. The core goal of the adaptive prior is to minimize the Kullback-Leibler (KL) divergence between the variational posterior and the prior, thereby maximizing the Evidence Lower Bound (ELBO).

Proposition 1 (ELBO Improvement by Adaptive Prior). Let $\mathcal{L}_{\text{stat}}(\mathbf{x})$ and $\mathcal{L}_{\text{adapt}}(\mathbf{x})$ denote the ELBO values under a static prior $p_{\text{stat}}(\mathbf{z}, k)$ and our adaptive prior $p_{\text{adapt}}(\mathbf{z}, k|S, T)$, respectively. Assuming that the conditional prior $p(\mathbf{z}|k)$ is consistent across both models, the adaptive prior will tighten the ELBO:

$$\mathcal{L}_{\text{adapt}}(\mathbf{x}) \geq \mathcal{L}_{\text{stat}}(\mathbf{x}). \tag{16}$$

Proof sketch. The ELBO $\mathcal{L}$ can be factored into a reconstruction term ($\mathcal{R}$), a latent KL term ($\mathcal{D}_{\mathbf{z}}$), and a cluster KL term ($\mathcal{D}_k$):

$$\mathcal{L} = \mathcal{R} - \mathcal{D}_k - \mathcal{D}_{\mathbf{z}}. \tag{17}$$

Where the cluster KL term is $\mathcal{D}_k = \mathrm{D}_{\mathrm{KL}}(q_\phi(k|\mathbf{x})\|p(k))$ (details in Appendix A.1).

The difference in ELBO, $\Delta\mathcal{L} = \mathcal{L}_{\text{adapt}} - \mathcal{L}_{\text{stat}}$, is primarily driven by the change in the cluster KL term: $\Delta\mathcal{L} \approx \mathcal{D}_k^{\text{stat}} - \mathcal{D}_k^{\text{adapt}} = \mathrm{D}_{\mathrm{KL}}(q_\phi(k|\mathbf{x})\|p_{\text{stat}}(k)) - \mathrm{D}_{\mathrm{KL}}(q_\phi(k|\mathbf{x})\|p_{\text{adapt}}(k|S, T))$. The adaptive prior $p_{\text{adapt}}(k|S, T)$ is designed to be conditioned on the input motion context (S, T), allowing it to better approximate the variational posterior $q_\phi(k|\mathbf{x})$ than the fixed, static prior $p_{\text{stat}}(k)$. Since the KL divergence is minimized when the prior matches the posterior, we have:

$$\mathrm{D}_{\mathrm{KL}}(q_\phi(k|\mathbf{x})\|p_{\text{adapt}}(k|S, T)) \leq \mathrm{D}_{\mathrm{KL}}(q_\phi(k|\mathbf{x})\|p_{\text{stat}}(k)). \tag{18}$$

Consequently, $\Delta\mathcal{L} \geq 0$. The adaptive prior tightens the ELBO by dynamically reducing the penalization for divergence between the inferred semantics and the prior semantics.

Corollary 1 (Improved Latent Separability). The act of reducing the cluster KL divergence $\mathrm{D}_{\mathrm{KL}}(q(k|\mathbf{x})\|p(k))$ forces the mixture weights $p(k)$ to track the inferred semantic assignments $q(k|\mathbf{x})$ more closely. This leads to sharper and less ambiguous assignments, which correspond to tighter clustering in the latent space. Theoretically, this reduction implies lower within-cluster variance and higher between-cluster variance in the latent space, confirming improved semantic disentanglement. This gain is empirically measured by higher Calinski–Harabasz and lower Davies–Bouldin scores.

Regularization for Stability. To constrain the powerful adaptivity of $p_{\text{adapt}}(k|S, T)$ and prevent the mixture weights $\pi(\cdot, t) = p_{\text{adapt}}(k|S, T)$ from overfitting to noisy temporal dynamics, we impose dual regularization terms: a temporal-smoothness term and an entropy term:

$$\mathcal{R}_{\text{adapt}} = \lambda_{\text{temp}} \sum_t \|\pi(\cdot, t) - \pi(\cdot, t\text{–}1)\|^2 - \lambda_{\text{ent}} \sum_t H(\pi(\cdot, t)). \tag{19}$$

The first term ensures **Lipschitz continuity over time** by bounding the change in mixture weights between adjacent frames. The second term, the negative entropy of the mixture weights, promotes sufficient mixture entropy and prevents any single component from completely dominating the mixture, thus maintaining diversity.

Empirical Validation. As demonstrated, the theoretical advantages of the adaptive prior are confirmed quantitatively by reporting: (i) The observed improvement $\Delta\mathrm{ELBO}$ over the static baseline, (ii) The average per-frame $\mathrm{KL}(q(k|\mathbf{x}_t)\|p(k))$ reduction, and (iii) The corresponding improvement in clustering metrics (CH/DB scores). These results validate that the adaptive prior learns a more expressive and structurally coherent motion manifold.

2. Appendix B

The Hierarchical Cross-Modal Attention (HCA) mechanism is designed to capture fine-grained correspondences between linguistic semantics and motion kinematics at multiple granularities.

Input Representation and Feature Projection Given the input motion features $\mathbf{z}^k \in \mathbb{R}^{T\times J\times D}$ (where T is temporal length, J is joint count, and D is feature dimension), we first flatten the spatial-temporal dimensions to obtain $\mathbf{z}^k_{\text{flat}} \in \mathbb{R}^{(T\cdot J)\times D}$. The text embeddings are processed at two levels: textbfWord-level features: $\mathbf{c}_{el} \in \mathbb{R}^{L\times D}$ obtained from CLIP text encoder with learnable projection. **Sentence-level feature**: $\mathbf{c}_s \in \mathbb{R}^D$ computed via temporal averaging of word embeddings. The temporally-grounded embedding $\mathbf{c}_{\text{eg}} \in \mathbb{R}^{T\times D}$ is generated as:

$$\mathbf{c}_{\text{eg}} = \mathbf{c}_{\text{s}} \oplus \text{MLP}(\mathbf{P}_{1:T}), \tag{20}$$

where $\mathbf{P}_{1:T} = \text{PositionalEncoding}(1 : T) \in \mathbb{R}^{T\times D}$, $\oplus$ denotes broadcast addition of the sentence vector $\mathbf{c}_s$ to all T rows, and the MLP consists of two linear layers with ReLU activation.

Hierarchical Cross Attention. In most denoisers, the CLIP-based text embedding is applied globally to the motion latent, which often fails to capture fine-grained correspondences between specific joints or frames and individual words in the text. Flat attention mechanisms often fail to align complex sentence structures with hierarchical motion semantics. We propose HCA, a two-level alignment mechanism integrating local (word-level), global (sentence-level) and Motion-Text cross-attention reasoning:

Formally, given motion features $z^{\text{k}}_{\text{t}} \in \mathbb{R}^{T\times J\times D}$ (flattened to $\mathbf{z}^k_{\text{flat}} \in \mathbb{R}^{(T\cdot J)\times D}$), word-level text embeddings

$c_{\text{el}} \in \mathbb{R}^{L\times D}$, and sentence-level embedding $c_{\text{s}} \in \mathbb{R}^{D}$, we compute projection matrices:

$$Q_{\text{CLC}} = c_{\text{cond}} W_{Q_{\text{CLC}}}, \qquad Q_{\text{M}} = z^{\text{k}}_{\text{flat}} W_{Q_{\text{M}}}, \tag{21}$$

$$K_{\text{el}} = c_{\text{el}} W_{K_{\text{EL}}}, \qquad K_{\text{eg}} = c_{\text{eg}} W_{K_{\text{EG}}}, \tag{22}$$

$$V_{\text{el}} = c_{\text{el}} W_{V_{\text{EL}}}, \qquad V_{\text{eg}} = c_{\text{eg}} W_{V_{\text{EG}}}. \tag{23}$$

The linguistic embedding c_{el} is obtained by feeding the raw word-level feature c_{w} into CLIP's text encoder and subsequently projecting the output with a learned linear layer. A sentence-level feature c_{s} is produced by averaging c_{w} along the temporal dimension. The temporally-grounded embedding c_{eg} is generated by adding to c_{s} a time-dependent signal that is first position-encoded and then transformed by a MLP.

$$\text{A}_{\text{CLC}} = \text{softmax}\Big(\frac{Q_{\text{CLC}} K^{\top}_{\text{EL}}}{\sqrt{d}}\Big) V_{\text{EL}}, \tag{24}$$

$$\text{A}_{\text{MWC}} = \text{softmax}\Big(\frac{Q_{\text{M}} K^{\top}_{\text{EG}}}{\sqrt{d}}\Big) V_{\text{EG}}. \tag{25}$$

where d is the attention dimensionality. This hierarchical design enables the denoiser to precisely align motion with text at both fine-grained and global semantic levels, improving motion-text fidelity and detail.

Dual-Attention Mechanism HCA employs two parallel attention computations: Conditioned Language Context (CLC) Attention: $\mathbf{Q}_{CLC} \in \mathbb{R}^{1\times d}, \mathbf{K}_{EL} \in \mathbb{R}^{L\times d}, \mathbf{V}_{EL} \in \mathbb{R}^{L\times d}, \mathbf{A}_{\text{CLC}} \in \mathbb{R}^{1\times d}$. Motion-Word Context (MWC) Attention: $\mathbf{Q}_{M} \in \mathbb{R}^{(T\cdot J)\times d}, \mathbf{K}_{EG} \in \mathbb{R}^{T\times d}, \mathbf{V}_{EG} \in \mathbb{R}^{T\times d}, \mathbf{A}_{\text{MWC}} \in \mathbb{R}^{(T\cdot J)\times d}$.

Implementation Details and Hyperparameters Attention dimension is $d = 256$ (consistent with base feature dimension D). Number of attention heads is 8 (yielding head dimension $d_h = 32$). MLP configuration is two linear layers with dimensions $[D \to D \to D]$, ReLU activation. Positional encoding is Sine-cosine encoding as in original Transformer. Normalization is LayerNorm applied before attention and after fusion. Multi-Head Attention Formulation: For each attention head i: $\text{Head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}^i_Q, \mathbf{K}\mathbf{W}^i_K, \mathbf{V}\mathbf{W}^i_V)$, $\text{MultiHead} = \text{Concat}(\text{Head}_1, \ldots, \text{Head}_h)\mathbf{W}_O$, where $\mathbf{W}^i_Q, \mathbf{W}^i_K, \mathbf{W}^i_V \in \mathbb{R}^{D\times d_h}$ and $\mathbf{W}_O \in \mathbb{R}^{h\cdot d_h \times D}$.

Computational Complexity Analysis The computational complexity of HCA is dominated by the attention operations: aBOUT CLC Attention, $\mathcal{O}(L \cdot d)$ where L is text sequence length ($L \le 77$). And MWC Attention is $\mathcal{O}(T^2 \cdot J \cdot d)$ where $T \cdot J$ represents the motion token count. In practice, with $T = 64$, $J = 22$, $L = 77$, and $d = 256$, the MWC attention constitutes approximately 85% of the total computational cost, making it suitable for real-time applications.

Algorithm 1 Hierarchical Cross-Modal Attention (HCA)

Require: Motion features $\mathbf{z}^k \in \mathbb{R}^{T\times J\times D}$, word embeddings $\mathbf{c}_{el} \in \mathbb{R}^{L\times D}$, sentence embedding $\mathbf{c}_s \in \mathbb{R}^{D}$
Ensure: Enhanced motion features $\mathbf{z}_{\text{enhanced}} \in \mathbb{R}^{T\times J\times D}$

// Step 1: Input projection and reshaping
$\mathbf{z}^k_{\text{flat}} \leftarrow \text{reshape}(\mathbf{z}^k, (T \cdot J, D))$
$\mathbf{c}_{cond} \leftarrow \text{LayerNorm}(\mathbf{c}_s)$
// Step 2: Generate temporally-grounded text features
$\mathbf{p}_t \leftarrow \text{PositionalEncoding}(1:T)\ \{T \times D\}$
$\mathbf{c}_{eg} \leftarrow \mathbf{c}_s + \text{MLP}(\mathbf{p}_t)\ \{\text{Broadcast addition}\}$
// Step 3: Compute dual attention pathways
$\mathbf{A}_{\text{CLC}} \leftarrow \text{MultiHeadAttention}(\mathbf{c}_{cond}, \mathbf{c}_{el}, \mathbf{c}_{el})$
$\mathbf{A}_{\text{MWC}} \leftarrow \text{MultiHeadAttention}(\mathbf{z}^k_{\text{flat}}, \mathbf{c}_{eg}, \mathbf{c}_{eg})$
// Step 4: Feature fusion and residual connection
$\mathbf{A}^{\text{broadcast}}_{\text{CLC}} \leftarrow \text{repeat}(\mathbf{A}_{\text{CLC}}, (T \cdot J, 1))$
$\mathbf{z}_{\text{fused}} \leftarrow \text{LayerNorm}(\mathbf{A}_{\text{MWC}} + \mathbf{A}^{\text{broadcast}}_{\text{CLC}})$
$\mathbf{z}_{\text{enhanced}} \leftarrow \text{reshape}(\mathbf{z}_{\text{fused}} + \mathbf{z}^k_{\text{flat}}, (T, J, D))$
return $\mathbf{z}_{\text{enhanced}}$

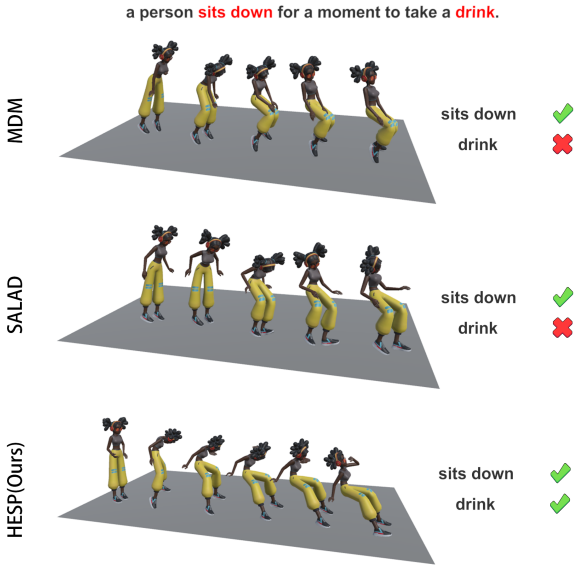


Figure 1. A visual comparison of MDM, SALAD, and our HESP.

3. Appendix C.Qualitative Analysis

3.1. Appendix C.1 Visual Samples

Qualitative Comparison across Models. Figure 1 presents a visual comparison among MDM [3], SALAD [2], and our HESP framework under the same text prompt: *"a person sits down for a moment to take a drink."* Each row depicts the motion sequence sampled from the respective model. As observed, MDM tends to generate coarse tran-

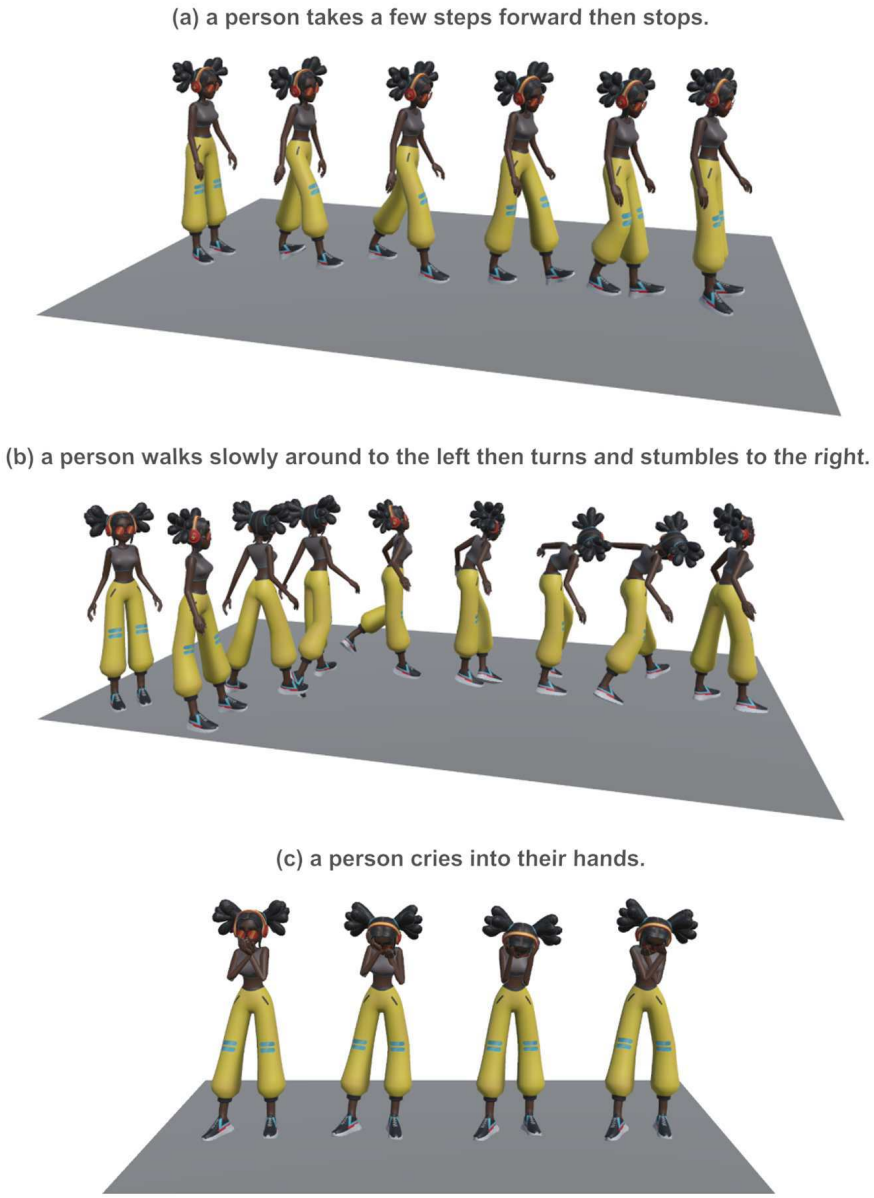


Figure 2. More visualizations of our HESP model.

sitions the person sits but no lifts the hand, often resulting in unrealistic. SALAD produces smoother trajectories but struggles to synchronize the hand implied by the text. In contrast, our HESP model generates a semantically coherent sequence in which the subject naturally bends the knees, lowers the torso, and performs a clear drinking gesture with consistent temporal dynamics. This demonstrates HESP's enhanced ability to capture fine-grained text–motion correspondences.

Diverse Motion Generation by HESP. To further illustrate the generative diversity and semantic grounding of HESP, Figure 2 showcases three representative motion samples generated under different textual descriptions:

- **(a)** *"a person takes a few steps forward then stops."* The model captures precise locomotion dynamics and deceleration at the stop phase.
- **(b)** *"a person walks slowly around to the left then turns and stumbles to the right."* The generated motion exhibits

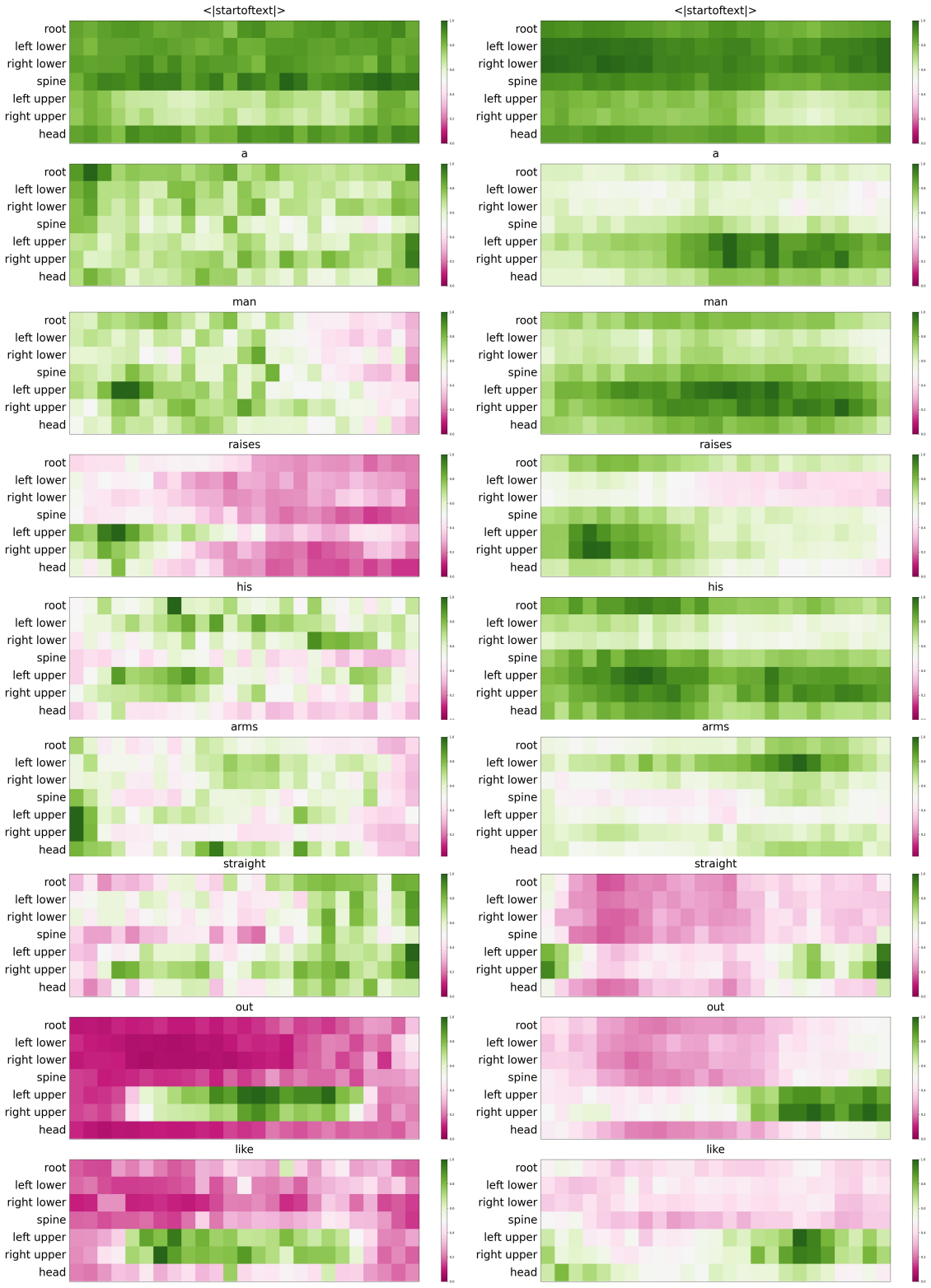


Figure 3. Attention heatmaps (word–joint alignment). Text description is “a man raises his arms straight out like a ’T’.” Under the same text description, the left column is the attention heatmap of the SALAD model, and the right column is our attention heatmap. We place the remaining words in Figure 4.

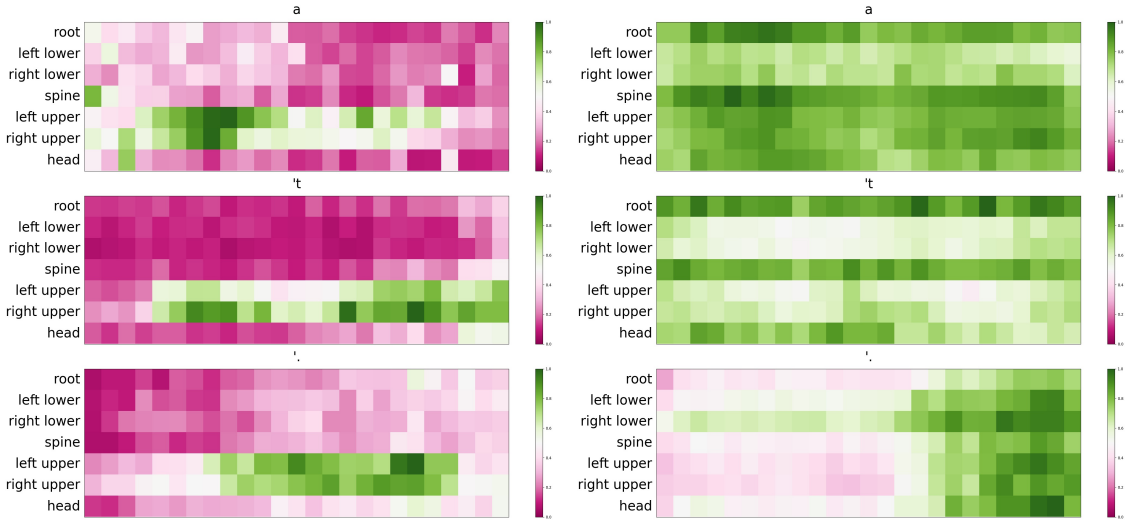


Figure 4. Attention heatmaps (word–joint alignment). Text description is "a man raises his arms straight out like a 'T'."

coherent multi-stage transitions with balanced body rotation and a realistic stumble.

- **(c)** *"a person cries into their hands."* HESP shows an expressive upper-body gesture with both hands covering the face.

These results demonstrate that HESP can generate motions that are not only physically plausible but also semantically faithful to fine-grained textual cues, supporting our claim of hierarchical semantic disentanglement and cross-modal interpretability.

3.2. Appendix C.2 Word Joint Attention Alignment

Objective. To further interpret how the proposed Hierarchical Cross-Modal Attention (HCA) enhances text–motion alignment, we visualize the attention weights between textual tokens and motion joints for both the baseline model (SALAD) and our HESP framework.

Example. Figure 3 and Figure 4 illustrates a case with the text prompt: *"a man raises his arms straight out like a 'T'."* The left column shows the attention heatmap of the SALAD model, while the right column corresponds to our HESP model under the same text description. We adopt a **watermelon-inspired colormap** in which greenish tones indicate higher attention weights and pinkish tones denote lower weights. This palette provides a smooth perceptual gradient and makes the activation pattern more interpretable.

Observations. In SALAD, the attention distribution is diffuse and often dominated by the torso joints, failing to more focus explicitly on the right regions. In contrast, our model exhibits sharply localized attention on the joints when attending to the key tokens *"raises"* and *"arms"*, indicating a clear correspondence between linguistic semantics and the associated body parts.

Discussion. This qualitative evidence supports the claim that HCA learns structured, interpretable cross-modal relationships: word-level attention concentrates on motion-relevant joints, while sentence-level attention captures holistic body configurations. Additional examples for other text descriptions are provided in Figure 11, Figure 12, Figure 9, Figure 10. Such structured attention patterns emerge without any explicit supervision, demonstrating the model's capability for self-aligned semantic grounding. Figure 1 qualitative results complement the qualitative results in Figure 11 and provide visual evidence that HESP learning links textual semantics to joints, thus confirming the effectiveness and interpretability of the proposed hierarchical framework.

3.3. Appendix C.3 Long-sequence Motion Generation

Objective. To demonstrate the capability of HESP in handling long-term, text-controlled motion generation, we extended our model to synthesize continuous single-person sequences that consist of multiple semantically distinct action segments. The complete visualizations are shown in Figures 5 and 6.

Setup. The long sequence contains three consecutive text descriptions:

(a) *A man walks forward while swinging his shoulders from side to side.*

(b) *A person holds their arms in the air and dances around.*

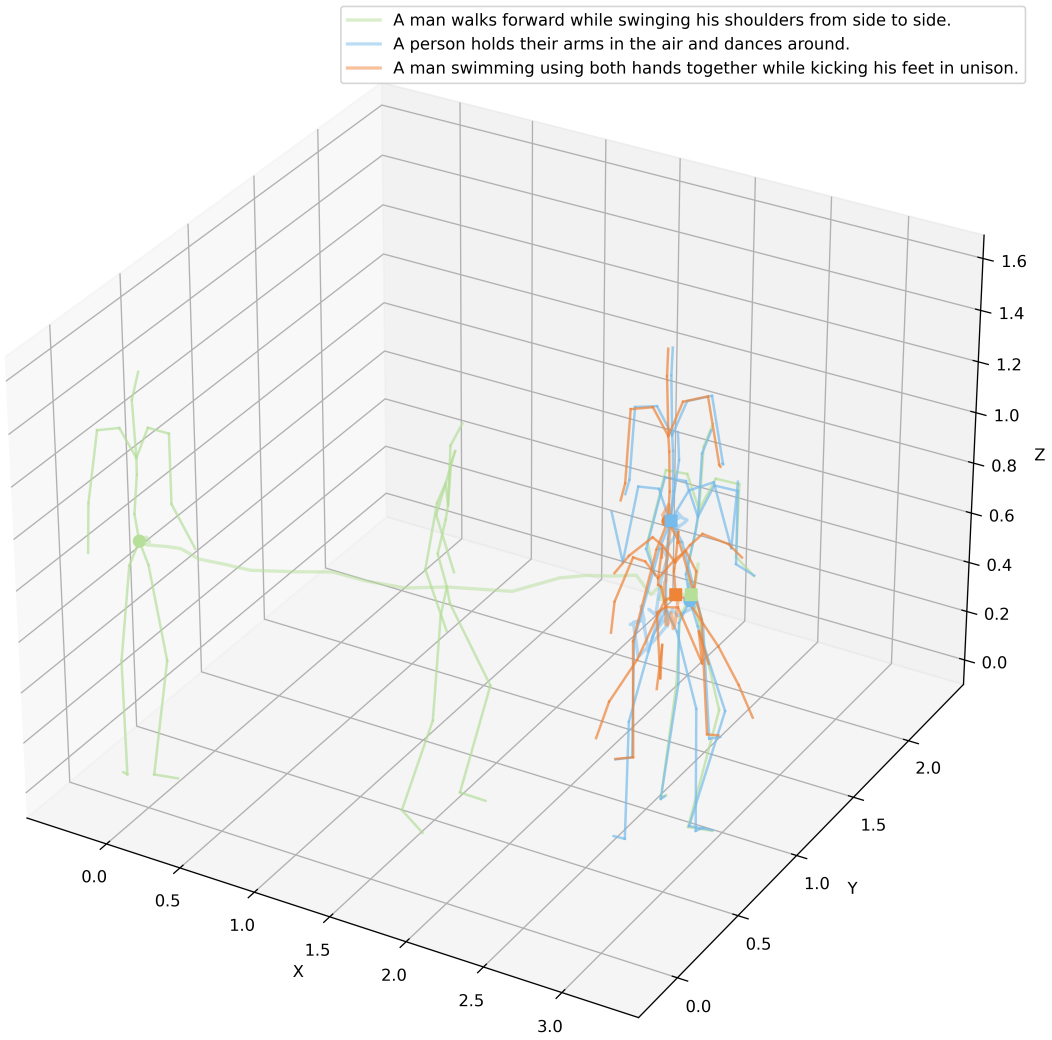


Figure 5. **Trajectory overlay with key poses for long-sequence generation.** Colored lines represent the root-joint trajectories for each textual segment, with circles and squares marking segment start and end points, respectively. Key skeletons depict intermediate poses sampled from each segment. HESP produces smooth spatial transitions and coherent cross-action dynamics in a single continuous sequence.

(c) *A man swimming using both hands together while kicking his feet in unison.*

Each segment lasts 196 frames, and the transitions between segments are generated automatically without manual temporal segmentation or stitching.

Pose Grid Overview. Figure 6 provides a frame-by-frame visual overview of the entire generated sequence. Every 10 frames are sampled and arranged into a grid. Colors indicate the semantic segment boundaries, where each color corresponds to one text description. The figure shows that HESP produces temporally coherent and semantically consistent motions across all three segments: walking $\rightarrow$ dancing $\rightarrow$ swimming. Smooth transitions and consistent body configurations demonstrate the model's ability to maintain structural continuity throughout extended sequences.

Visualization 2: Trajectory Overlay with Key Poses. To further analyze global motion continuity and spatial coherence, Figure 5 overlays the root trajectories and repre-

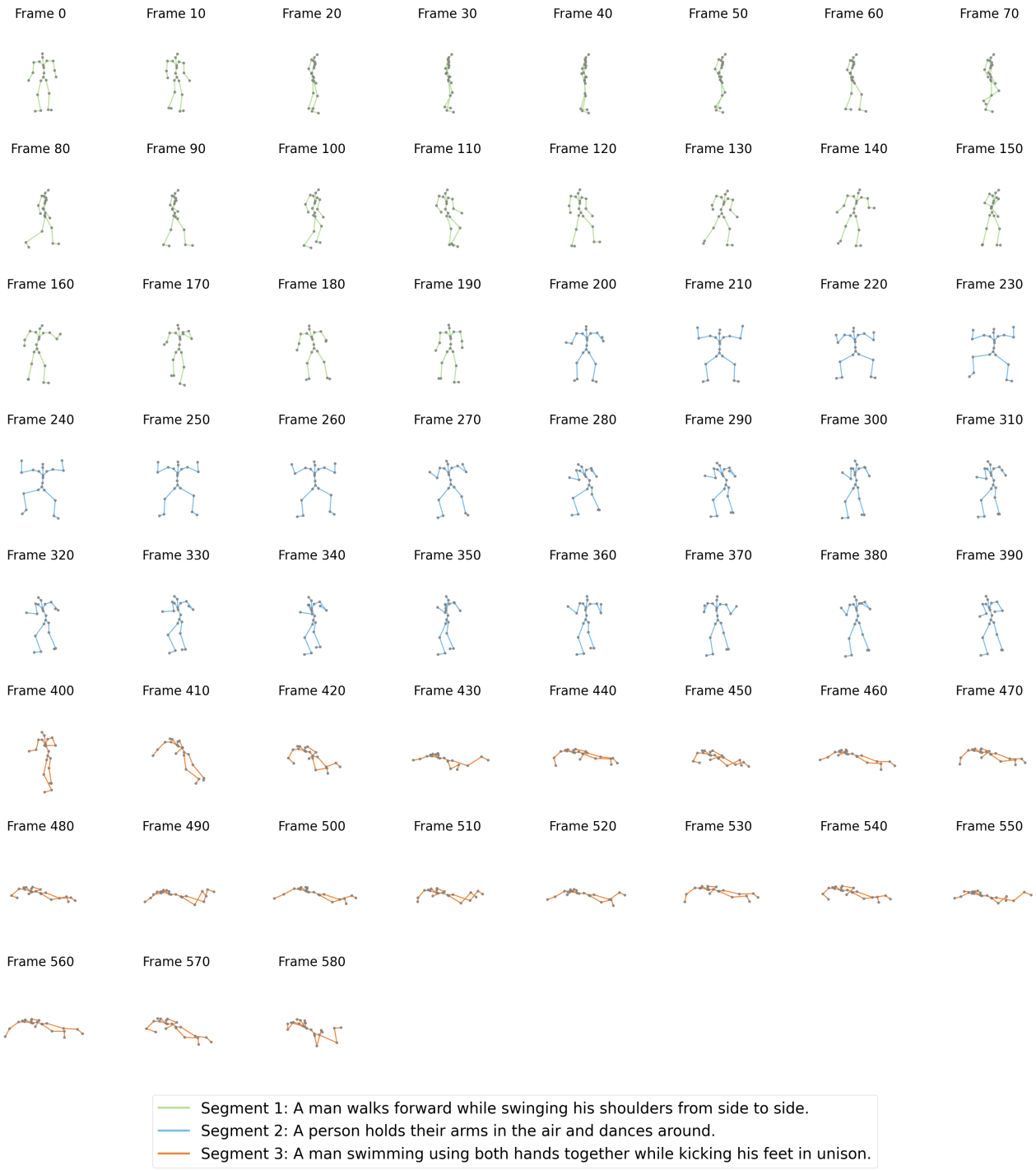
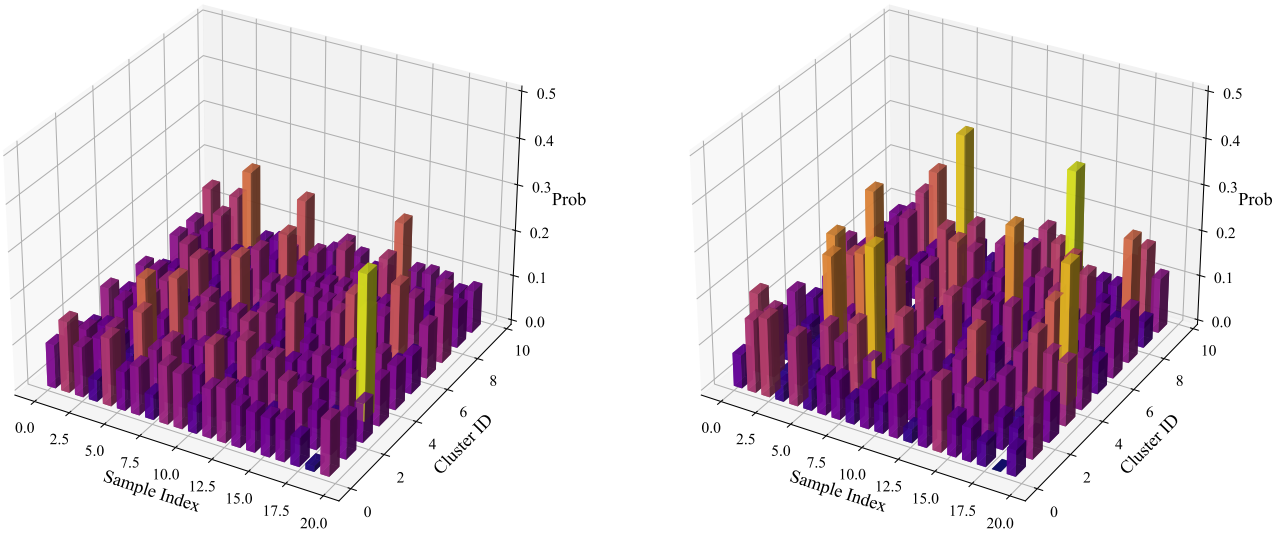


Figure 6. **Long-sequence motion generation (pose grid visualization).** Each colored segment corresponds to a different textual prompt: (a) A man walks forward while swinging his shoulders from side to side. (b) A person holds their arms in the air and dances around. and (c) A man swimming using both hands together while kicking his feet in unison. Frames are sampled every 10 time steps and arranged in a grid. HESP maintains semantic fidelity and temporal smoothness across transitions without manual stitching.

sentative key poses of the same long sequence. Each colored trajectory corresponds to one textual segment. Start and end positions are marked with circles and squares respectively. Key skeletal frames illustrate how the model



(a) Standard VAE (Baseline). (b) AG-VAE.

Figure 7. Comparative analysis of cluster probability distributions.

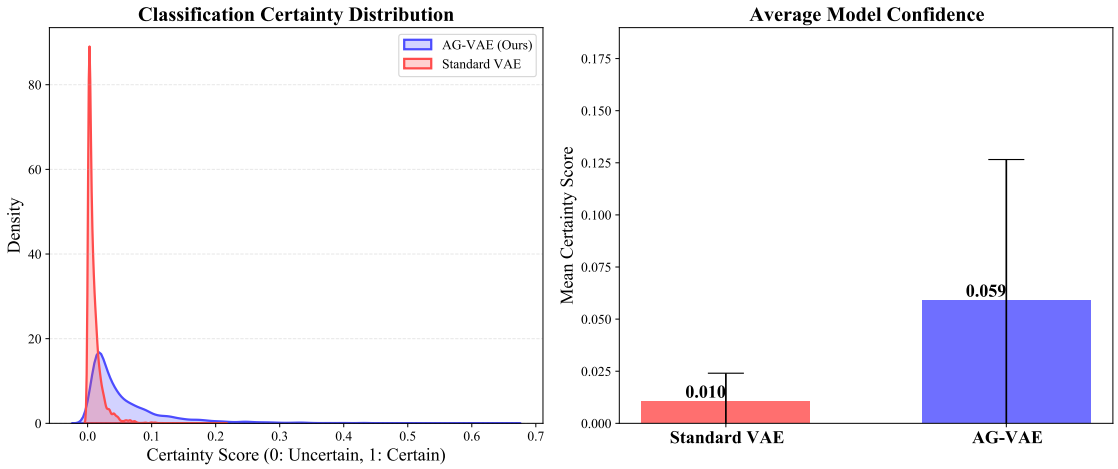


Figure 8. Quantitative Comparison of Classification Certainty.

transitions from one activity to the next while preserving realistic body posture.

The overlay clearly shows that the walking segment advances forward, the dancing segment explores a localized spatial region, and the swimming segment shifts to a horizontal, rhythmic motion pattern. These results confirm that the model's temporal dynamics and semantic transitions are both continuous and physically plausible.

4. Appendix D

We demonstrate that the advantage of Figure 1-(b) over Figure 1-(a) lies in the transition from isotropic compression to structured manifold learning. The standard VAE (Figure 1-(a)) suffers from center collapse, where different motion semantics become entangled near the origin due to a unimodal Gaussian prior. In contrast, the AG-VAE (Figure 1-(b)) explicitly structures the latent space into multiple semantically coherent submanifolds.

As suggested, we present a baseline comparison of clustering probability distributions (see Figures 7a, 7b). The results (in Figure 8) show that AG-VAE achieves a much higher assignment confidence (mean ≈ 0.59) compared to the standard VAE (≈ 0.01), providing a more deterministic and disentangled prior for the subsequent diffusion generation.

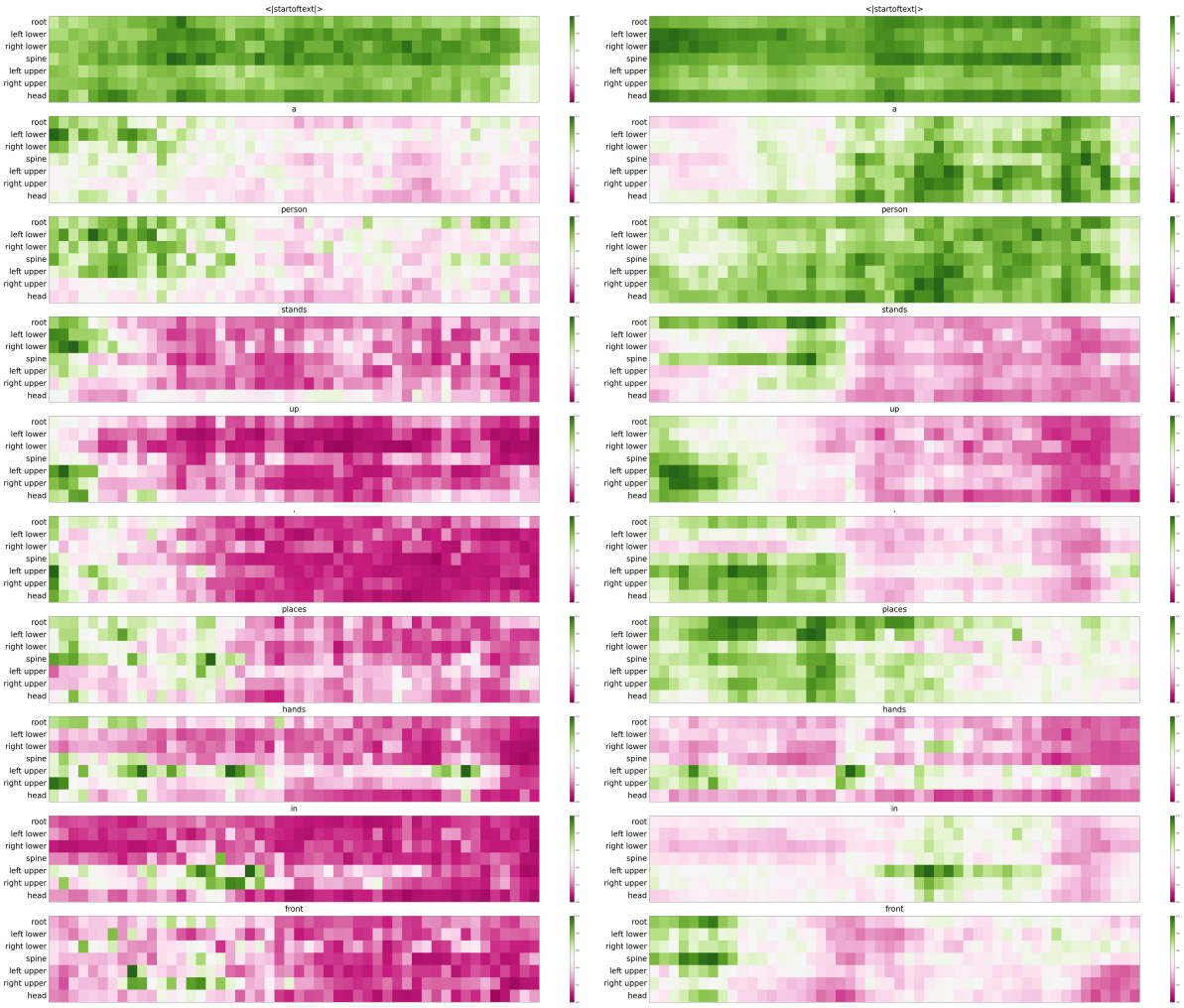


Figure 9. Attention heatmaps (word–joint alignment). Text description is “a person stands up, places hands in front of chest and down, and walks in a counterclockwise circle and sits down.” Under the same text description, the left column is the attention heatmap of the SALAD model, and the right column is our attention heatmap. We place the remaining words in Figure 10.

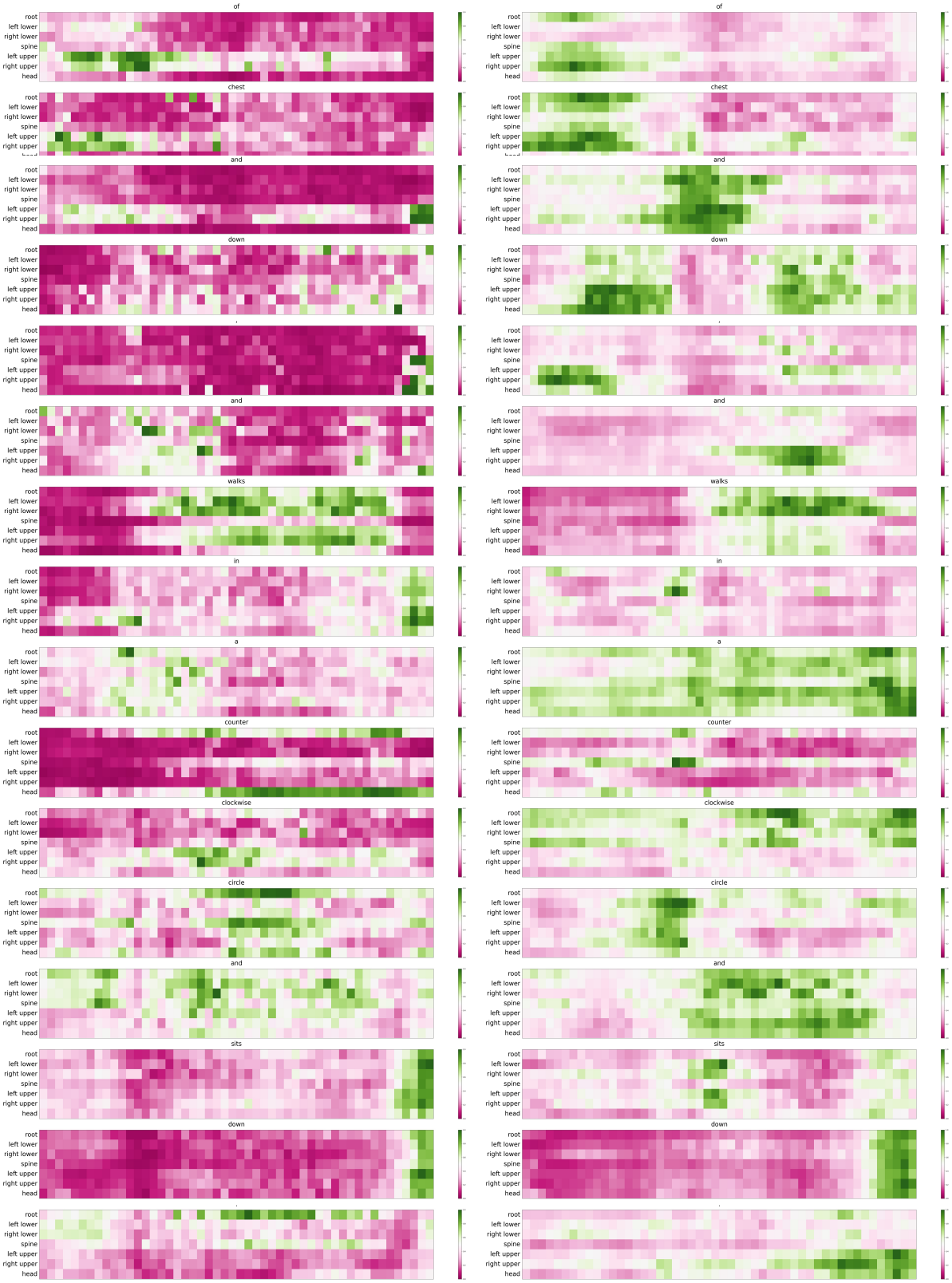


Figure 10. Attention heatmaps (word–joint alignment). Text description is “a person stands up, places hands in front of chest and down, and walks in a counterclockwise circle and sits down.”

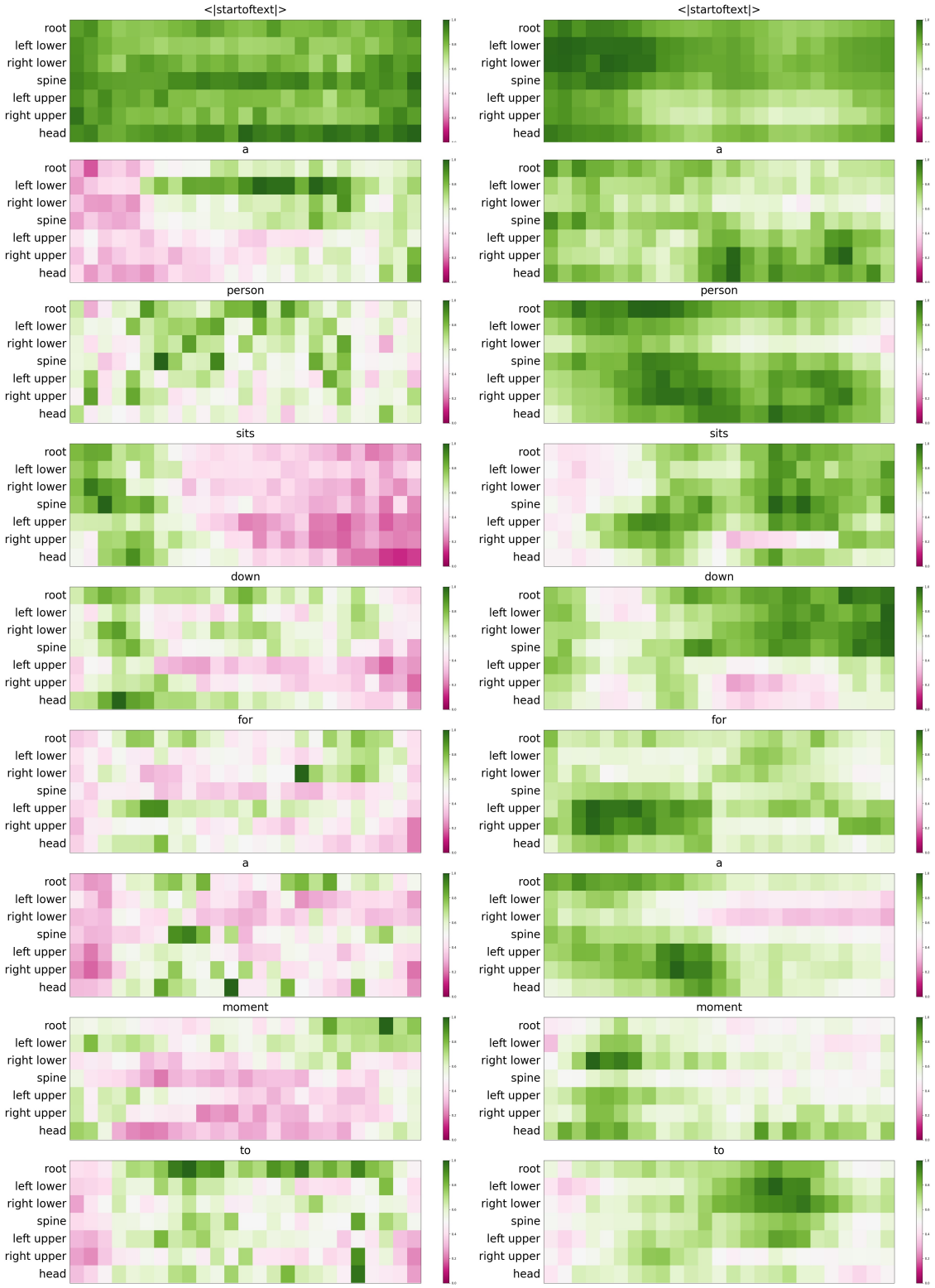


Figure 11. Attention heatmaps (word–joint alignment). Text description is “a person sits down for a moment to take a drink.” Under the same text description, the left column is the attention heatmap of the SALAD model, and the right column is our attention heatmap. We place the remaining words in Figure 12.

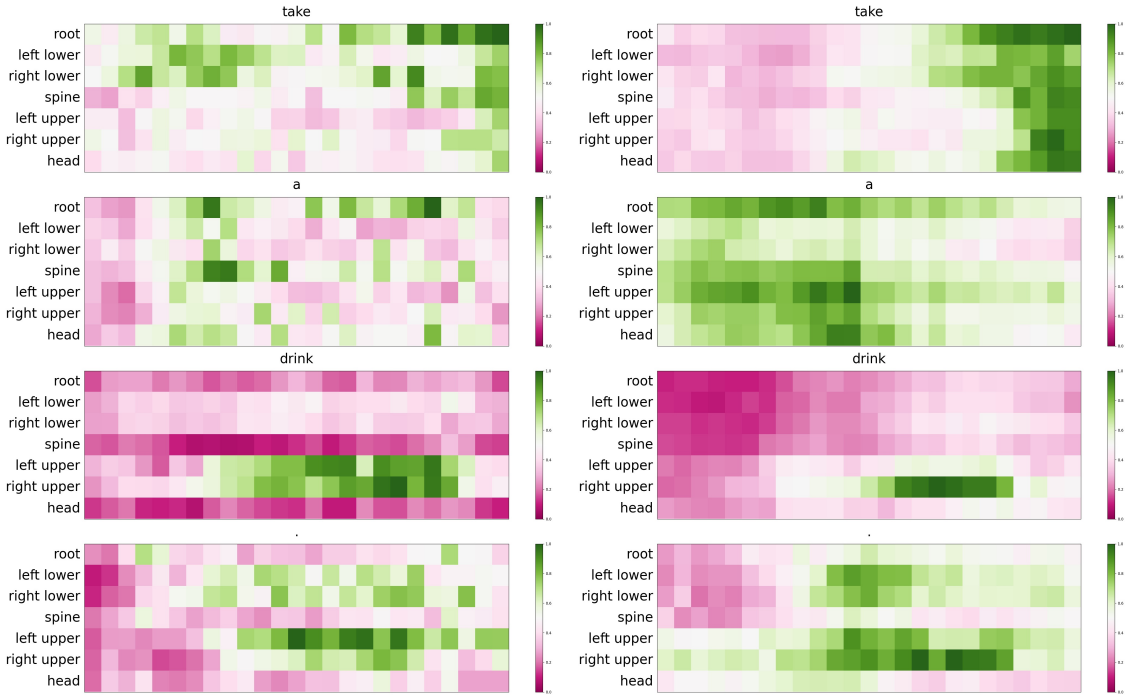


Figure 12. Attention heatmaps (word–joint alignment). Text description is “a person sits down for a moment to take a drink.”

References

[1] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648, 2016. 1

[2] Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 7158–7168, 2025. 5

[3] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In The Eleventh International Conference on Learning Representations, 2023. 5