BRING SCRIPTS TO ACTING! CONTENT-VARIANT CONTROL FOR EMOTION-ALIGNED ACTING GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce **Script-to-Acting**, which generates controllable acting from a raw script. Existing text-to-motion and motion-stylization methods typically keep the motion content fixed and only rescale style or intensity, a paradigm we term *content-consistent* control. Acting is different: when the requested emotional state changes, the content of the performance can change with it. As the intensity along an emotion axis increases, the acting may shift from a submissive plea to more confrontational actions, such as a waving reprimand, a shove-away, or a shoulder shake. We therefore construct **TheaterAct**, a multimodal dataset for acting generation containing 7.2 hours of professional performance from 59 characters, with aligned scripts, dialogue, motion, and fine-grained emotion annotations. Based on this dataset, we design **Acting Mind**, **Acting Morph**, and **Acting Creation**. Acting Mind uses a dramaturgical context projector to interpret script, character, and dialogue through structured relationship, motivation, and emotion cues. Acting Morph represents dramatic emotion as a token-wise acting-mode matrix whose coordinates are grounded in professional performances indexed by script semantics, emotion axis, and continuous intensity. Rather than broadcasting a global style code, the matrix allocates dramatic control to individual script tokens and produces an emotion-aware motion feature for synthesis. Acting Creation then realizes this *content-variant* control on top of any pretrained motion generator, upgrading it to support emotion-aligned acting without architectural changes. Experiments show that our method outperforms existing motion-generation models in script alignment, emotion-intensity alignment, physical plausibility, and *content-variant* acting. Notably, the proposed modules can be integrated with multiple motion-generation backbones as a generic enhancement.

1 INTRODUCTION

Acting is a remarkable form of human intelligence. It transforms narrative intent, emotion, character, and social relationships into meaningful bodily behaviour that gives a scene its dramatic force and emotional resonance. For animation, film, games, virtual reality, and intelligent robotics, generating expressive acting rather than mere motion is crucial for creating coherent and lifelike behavior. Current AI systems cannot yet reproduce the human intelligence required for genuine acting. Bridging this gap requires grounding abstract linguistic and artistic intent in bodily motion, finely controlling emotional states, and jointly aligning the performance with character, relational, and physical constraints. We identify three fundamental challenges that current methods must overcome.

First, scripts do not speak the language of motion. Existing text-to-motion methods take explicit action descriptions such as "walks forward" or "raise the right arm" and align motion to these literal instructions (Tevet et al., 2022; Jiang et al., 2023). A theatrical script instead provides dialogue, plot, relationships, and implied intent, while leaving the physical performance unstated. The line "To be, or not to be" does not specify whether the speaker should kneel, pace, recoil, or raise a hand. Consequently, existing language-to-motion systems can generate fluent motion that does not function as a performance. We address this problem with Acting Mind, a Dramaturgical Context Projector that uses professional theatre knowledge to map character attributes and relationships,

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.

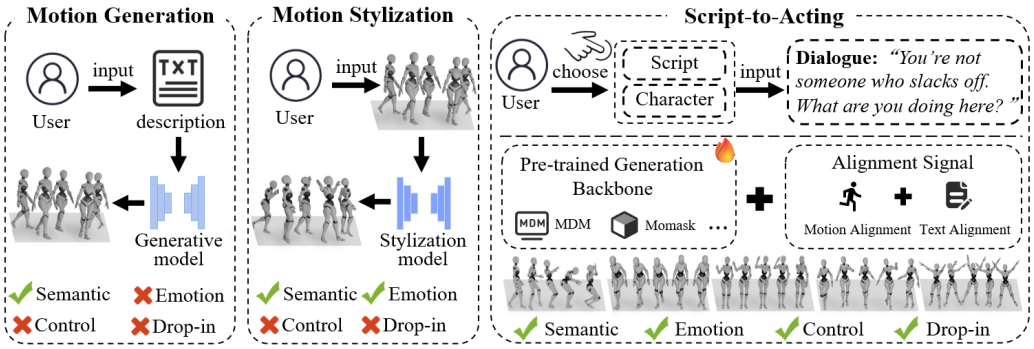

Figure 1: Motion generation and motion stylization either follow a caption or restyle an existing clip, whereas Script-to-Acting, given a script, a character, and a line, generates content-variant acting on a pretrained backbone. Emotion and control are enabled as a drop-in alignment signal, so both the strength and the motion identity of the performance change.

motivation, and dialogue into a structured acting context. This context is not a motion caption; it provides information needed for later acting-mode selection and motion generation.

Second, emotion reshapes action, yet existing methods usually keep the action fixed. Motion stylization and emotion-conditioned generation typically preserve motion content while modifying style or intensity (Aberman et al., 2020; Kim et al., 2024; Guo et al., 2025b; Yu et al., 2024). We call this *content-consistent* control. Acting requires a different capability: as dramatic state changes, a performance may move between semantically distinct acting modes rather than merely rescaling one gesture. The model must therefore support *continuous, content-variant* control—a trajectory along a dramatic axis on which the realized motion can change its acting motif. We address this challenge with **Acting Morph**. Existing methods commonly encode emotion or style as a single global latent code that is broadcast across the sequence; changing its strength therefore primarily changes motion amplitude, speed, or energy while leaving the underlying acting mode fixed. In contrast, Acting Morph represents the requested dramatic state as a token-wise acting-mode matrix. Its coordinates are grounded by professional TheaterAct performances indexed jointly by script semantics, the active emotion axis, and continuous intensity. Rather than treating a retrieved clip as the output or broadcasting one global style vector, Acting Morph allocates dramatic control to individual script tokens and binds the resulting professional motif coordinates to the evolving motion sequence, producing an emotion-aware motion feature for synthesis. Consequently, varying intensity can traverse distinct acting motifs along a dramatic axis, so intensity controls which acting mode is realized, not only how strongly a fixed action is performed.

Third, acting requires jointly aligning multiple interacting factors. The appropriate motion for a dialogue depends on character attributes and relationships, motivation, and dramatic emotions. These factors are coupled rather than independent, but existing general-purpose text-to-motion datasets provide limited supervision for their joint alignment. We address this data bottleneck with **TheaterAct**, a multimodal dataset of script–dialogue–acting alignment containing 7.2 hours of professional theatrical performance from 59 characters. It provides temporally aligned dialogue–motion pairs with structured annotations for character information and relationships, motivation, emotion, and continuous intensity along 6 emotion axes, enabling supervised training and evaluation of script-conditioned, emotion-aligned acting.

Towards this end, we propose **Script-to-Acting**, a unified framework that interprets a raw script, applies fine-grained dramatic control, and generates expressive acting under narrative, character, emotional, and physical constraints. The design is generic: **Acting Mind** provides structured script interpretation, **Acting Morph** provides transferable content-variant control, and **Acting Creation** synthesizes motion through a conditioned generative model. Across quantitative, qualitative, ablation, and human evaluations, Script-to-Acting improves script alignment, acting rationality, character consistency, and emotion expressiveness over existing motion-generation baselines, and can be integrated with multiple backbones to enhance emotion control. Our contributions are as follows:

- We formulate **Script-to-Acting**, a framework that turns scripts directly into expressive, physically plausible, contextually aligned acting, with continuous control over dramatic

intensity instead of discrete emotion labels. The modules can be integrated into existing motion-generation models to enhance emotion-aligned generation and fine-grained control.

- We identify **content-variant** control as the key capability that distinguishes acting from styled motion: as emotional intensity increases, the acting mode can change rather than remaining fixed. We realize this through **Acting Morph**, a lightweight module that enables intensity to traverse distinct acting motifs instead of merely rescaling a fixed action.
- We release **TheaterAct**, the first multimodal dataset centered on script–dialogue–acting alignment and fine-grained dramatic emotions, with script-level context and rich annotations. It enables supervised training and evaluation of script-conditioned, emotion-aligned acting, and supports continuous intensity control along multiple emotion axes, providing a benchmark substrate for future work on expressive acting generation.

2 Related Works

Motion Generation. Text-to-motion generation maps natural language descriptions to human motion sequences. Recent work has largely adopted denoising diffusion and discrete motion tokens to improve quality and semantic control (Tevet et al., 2022; Zhang et al., 2024; Guo et al., 2025a; Chen et al., 2023; Zhou et al., 2024; Jiang et al., 2023; Pinyoanuntapong et al., 2024; Guo et al., 2024). These methods remain a semantic-to-motion map: they do not read a performance script or generate acting aligned with plot dialogue. Closer to our setting, emotion-enriched generation injects a discrete feeling into a motion (Yu et al., 2024), and co-speech gesture drives frame-level movement from audio or dialogue (Li et al., 2021; Zhu et al., 2023). Script-to-Acting differs in three respects. First, emotion is typically a discrete label, with little control over continuous intensity. Second, even continuous knobs stay *content-consistent*: style or amplitude changes while the action stays fixed, whereas we allow *content-variant* acting as intensity changes. Third, existing methods align motion with only a single factor—typically an emotion tag or a local speech signal—whereas our approach jointly aligns character information and dramatic emotion.

Motion Stylization. Motion style transfer aims to modify the stylistic characteristics of a motion while preserving its semantic content. Recent diffusion-based and multimodal methods represent style as a global vector or style code that is applied uniformly across the sequence (Raab et al., 2024; Zhong et al., 2024; Song et al., 2024; Guo et al., 2025b). As a result, changing style or intensity mainly rescales amplitude or stylistic strength while the underlying action content remains fixed. In contrast, Script-to-Acting targets content-variant acting, where dramatic state changes can transform the action content itself. We design a new module, Acting Morph, which replaces the global style code with a token-wise motif-coordinate matrix, so that intensity selects different acting motifs per token instead of only rescaling a fixed motion.

Controllable Generation. Controllable motion generation conditions on text, style, music, action, audio, or skeletal trajectories (Guo et al., 2025a; Tseng et al., 2023; Guo et al., 2025b; Lee et al., 2023; Zhu et al., 2023; Huang et al., 2023; Lu et al., 2022). Diffusion-based methods align motions with text or music, and trajectory-based methods edit joint positions, but neither supports semantically meaningful control for dramatic performance. A separate line uses an LLM as a planner or choreographer (Wu et al., 2025; Wang et al., 2025). In contrast, our framework supports script-based conditioning with continuous emotion-intensity control along 6 dramatic axes, matching the granularity of theatrical performance, while letting the action motif change as intensity varies.

3 Method

Overview. Script-to-Acting takes a script, a user-chosen character, a line of dialogue, and a user-specified emotion axis with an intensity, and returns a motion sequence for that line. Figure 2 shows the three components. **Acting Mind** (Section 3.1) maps the script to a structured dramatic plan and a quantitative emotion vector. These outputs condition **Acting Morph** (Section 3.2), which retrieves a professional acting prior at the requested point in the acting-mode family. **Acting Creation** (Section 3.3) then synthesizes motion with a conditioned generative model that aligns three coupled factors: the plan, the retrieved prior, and the emotion target.

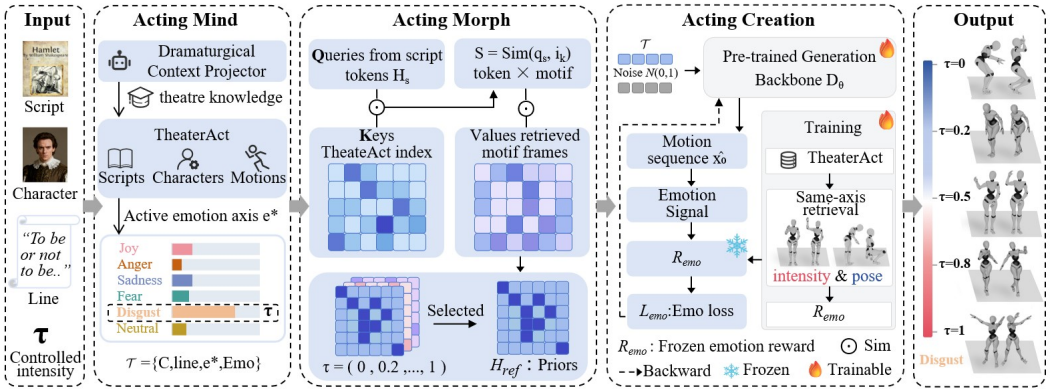

Figure 2: **Script-to-Acting.** Given a script, a character, a line, and an intensity τ along the active emotion axis $e^\star$, Acting Mind produces an acting plan $\mathcal{T}$. Acting Morph scores each script token against same-axis professional clips from TheaterAct, binding them into a token-wise acting-mode matrix S and encoding the selected columns as priors H_{ref}. Acting Creation then synthesizes motion conditioned on $\mathcal{T}$ and H_{ref}, yielding a content-variant family as τ is swept.

3.1 ACTING MIND: DRAMATURGICAL CONTEXT PROJECTION

Challenges. Scripts do not specify motion. Acting jointly aligns character, relationships, motivation, and given circumstances, whereas text-to-motion models rely on overly simplified action captions. Acting Mind therefore acts as a dramaturgical context projector and agent that converts a raw script into a structured acting plan.

Dramaturgical context projection. We address these challenges with Acting Mind, a dramaturgical context agent implemented by a projector Φ_{LLM}. Rather than treating a script line as a motion caption, the agent applies theatre knowledge about given circumstances, character C, relationships, motivation, and line-action cues, and organizes them into a structured acting context:

$$\Phi_{\text{LLM}}(T_{\text{script}}, C, \text{line}) \;\rightarrow\; \{C,\ \text{relation},\ \text{motivation},\ \text{line},\ \text{Emo}\}. \tag{1}$$

The output is neither a joint trajectory nor an action prompt. It is a plan for later acting-mode selection and motion generation. **Indexing the acting-mode family.** Each TheaterAct (Section 4) prior carries one active emotion axis and an intensity. We index the acting-mode family on 6 independent axes: Anger, Disgust, Fear, Joy, Sadness, and Neutral. The projector extracts an emotion vector from the script and character context; the user sets an active axis $e^\star$ and overwrites that coordinate with intensity $\tau \in [0, 1]$:

$$\text{Emo} = \text{EmotionExtract}(T_{\text{script}}, C), \quad \text{Emo}_{e^\star} = \tau, \quad \mathcal{T} = \{C,\ \text{line},\ e^\star,\ \text{Emo}\}, \tag{2}$$

where $\text{Emo} \in [0, 1]^6$ is consumed by Acting Morph and Acting Creation. With script, character, and line fixed, sweeping τ changes the requested point in the acting-mode family rather than substituting a new sentence. The plan $\mathcal{T}$ contains the dramatic context and requested emotional state, but it does not contain a demonstrated trajectory. Acting Morph uses $\mathcal{T}$ to retrieve professional priors by $(e^\star, \tau)$ and line semantics; Acting Creation synthesizes a new performance under the resulting conditions. This separation makes content-variant control possible: the controlled object is the acting-mode index, not a demonstrated motion.

3.2 ACTING MORPH: FROM GLOBAL STYLE TO TOKEN-WISE ACTING MODES

Motivation. Existing emotion-conditioned motion generation and motion stylization commonly represent emotion or style with a single global latent code. Such a code is broadcast across the sequence, so changing its strength mainly rescales the amplitude, speed, or energy of a fixed action. This content-consistent assumption is insufficient for acting: as dramatic intensity changes, the appropriate physical realization may move between distinct acting modes, such as kneeling prayer, one-hand rebuke, shove and turn, or charge and shake. We therefore replace the global style vector with a token-wise acting-mode matrix.

Token-wise acting-mode matrix. Given script-token features $H_s \in \mathbb{R}^{L_s \times d}$ and the dramatic state Emo from Acting Mind, Acting Morph builds a matrix $\mathbf{S}$ whose rows are script tokens and whose columns are professional motifs. Entry $\mathbf{S}_{n,k}$ is the binding strength of cue n to motif k under $(e^\star, \tau)$, so different tokens on the same line can bind to different acting modes rather than sharing one broadcast code.

Professional motif coordinates. For a query plan $\mathcal{T}$, we retrieve professional anchors consistent with line semantics, the active axis, and the requested intensity. A semantic embedding e^{sem} is obtained from the line with a frozen CLIP text encoder. For clip i, let $a_i \in \mathcal{A}$ and $\iota_i \in [0,1]$ denote its active axis and annotated intensity. We score it by:

$$\mathrm{Sim}(q,i) = \omega_{\mathrm{sem}} \cos\big(e_q^{\mathrm{sem}},\, e_i^{\mathrm{sem}}\big) + \omega_{\mathrm{emo}}\, \mathbf{1}[a_i = e^\star]\big(1 - |\iota_i - \tau|\big). \tag{3}$$

The first term preserves script alignment; the second indexes the requested point on the dramatic axis, where $\mathbf{1}[\cdot]$ is the indicator function. Main-table evaluation uses same-axis CLIP; the intensity residual $(1 - |\iota_i - \tau|)$ is used on the τ-grid. We then discard clips violating basic contact geometry. With world-coordinate floor $y = 0$, let $\mathrm{slide}(i)$ be the mean horizontal step of a contacting foot and $\mathrm{pen}(i)$ the ground penetration:

$$\mathrm{Fil}_{\mathrm{phy}}(i) = \mathbf{1}\big[\mathrm{slide}(i) \le \varepsilon_{\mathrm{s}} \ \wedge\ \mathrm{pen}(i) \le \varepsilon_{\mathrm{p}}\big]. \tag{4}$$

Clips with $\mathrm{Fil}_{\mathrm{phy}}(i) = 1$ become the motif columns of $\mathbf{S}$.

Local dramatic allocation. Retrieval supplies motif columns; it does not decide how each line-action cue is played. Acting Morph applies the same Sim at the token level to obtain $\mathbf{S}$ in Figure 2:

$$\mathbf{S}_{n,k} = \mathrm{Sim}(q_n, i_k), \qquad M_{\mathrm{ref}}(\tau) = \text{Top-}K(\mathbf{S}), \quad i_k \in \{i : \mathrm{Fil}_{\mathrm{phy}}(i) = 1\}. \tag{5}$$

Sweeping τ moves mass across columns, so intensity selects which acting mode is realized rather than how strongly a fixed action is performed. The selected columns are encoded as $H_{\mathrm{ref}} \in \mathbb{R}^{L_r \times d}$; the backbone applies its existing self-attention on $X = [H_{\mathrm{ref}}; H_m]$ to realize $\mathbf{S}$ (no new layer). Acting Creation then synthesizes motion under $\mathcal{T}$ and H_{ref}. At training time the retrieved motion is a prefix anchor for reconstruction; at inference the prefix remains visible while remaining frames are inpainted. Thus professional priors ground $\mathbf{S}$, but the full-body performance is generated rather than copied. Equation 3 scores a whole line against a clip, whereas Equation 5 scores each script token separately. In our ablation, a line-level variant uses the same Sim but queries the whole line via Equation 3, binding one motif to the entire line instead of individual tokens as in Equation 5.

3.3 ACTING CREATION: REALIZING MULTI-FACTOR ALIGNED ACTING

Challenges. (1) Acting must align the script, character, relationships, motivation, emotion axis, and continuous intensity. Acting Mind and Acting Morph encode these coupled controls at different levels, but they must be realized in one full-body performance. (2) The generated motion must remain natural and physically plausible while preserving these contextual and dramatic controls; reconstruction alone can produce fluent motion that ignores the requested acting mode.

Multi-factor realization. Acting Creation realizes the plan $\mathcal{T}$ from Acting Mind together with the motif coordinates H_{ref} from Acting Morph. These sources enter the generator through text conditioning from $\mathcal{T}$ (and the neighbor caption), motion tokens $X = [H_{\mathrm{ref}}; H_m]$ in the backbone, and a frozen emotion reward on the active axis. Intensity is not a scalar that rescales a global style; it has already selected the coordinates of $\mathbf{S}$ through τ, and it is written into $\mathrm{Emo}_{e^\star} = \tau$.

Multi-objective aligned realization. Let x_0 be a TheaterAct motion window (held-out test play excluded), x_t the noisy input at diffusion step t, and $\hat{x}_0$ the denoiser prediction of the clean motion. The reconstruction and coordinate terms are:

$$\mathcal{L}_{\mathrm{rec}} = \mathbb{E}\|\hat{x}_0 - x_0\|_2^2, \qquad \mathcal{L}_{\mathrm{ref}} = \mathbb{E}\|(\hat{x}_0 - M_{\mathrm{ref}}) \odot m_{\mathrm{ref}}\|_2^2, \tag{6}$$

where m_{ref} masks the bound motif coordinates. Dramatic alignment uses a frozen motion-to-emotion encoder φ, trained on non-held-out TheaterAct clips, which maps motion to a six-dimensional acting signal on $\mathcal{A}$:

$$\mathcal{L}_{\mathrm{emo}} = -\varphi_{e^\star}(\hat{x}_0). \tag{7}$$

The backbone D_θ is trained with:

$$\mathcal{L} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{ref}}\mathcal{L}_{\mathrm{ref}} + \gamma\mathcal{L}_{\mathrm{emo}}, \tag{8}$$

and gradients update only D_θ. The same φ is used for ρ and R_{emo} at evaluation. Copying a retrieved motion at the wrong intensity does not minimize $\mathcal{L}$: $\mathcal{L}_{\text{rec}}$ targets a new performance of the query line, and $\mathcal{L}_{\text{emo}}$ requires the active axis to be expressed in $\hat{x}_0$. The retrieved prefix is a professional anchor, not the output. We report physical plausibility at evaluation as sliding, floating, and ground penetration under the world-coordinate floor $y = 0$. Gains on these measures are consistent with grounding generation in professionally performed coordinates rather than unconstrained noise. The frozen R_{emo} can be inserted into a diffusion or VQ backbone without retraining φ.

4 DATASET: THEATERACT

We construct **TheaterAct**, a multimodal theatrical dataset for script-to-acting generation. *TheaterAct* is distinctive in three respects: script alignment, where each clip is bound to a play, a character C, and a line; theatrical expression, where motion comes from stage performance rather than everyday motion capture; and emotion alignment, where each clip is labelled with an active axis among six emotions and an intensity τ on that axis. Appendix A.3 gives statistics and field definitions.

(1) Collection. Clips are cropped from about 30 hours of theatrical source video. The release keeps 7.2 hours, 730 clips, at 20 fps, covering 10 plays and 59 character cards. **(2) Processing.** Human motion is extracted with GVHMR (Shen et al., 2024) and written in the 263-dimensional HumanML3D (Guo et al., 2022) format so that existing generators can consume it. Timestamps align each line with motion and with a motion caption. **(3) Annotation.** Labels sit in three blocks: play, character, and line–motion pair. The play block records plot, setting, props, and genre. The character block gives the given circumstances of C: a short description, personality, occupation, and a 10-way multi-label motivation. The line–motion block aligns a spoken line, 263-dimensional motion, and a HumanML3D-style caption, and marks the active emotion axis $e^\star$ with intensity τ.

5 EXPERIMENT

Setup. For each evaluation sample, the target emotion axis and intensity are obtained directly from its annotated test-set condition. Based on the selected axis and text semantics, our system retrieves professional acting priors from the TheaterAct library using CLIP similarity, which is strictly restricted to the same emotional axis. A separate protocol sweeps $\tau \in \{0, 0.2, \ldots, 1.0\}$ on the same lines for intensity control, unused in the main tables.

Configuration. Two datasets are employed: HumanML3D and TheaterAct. HumanML3D follows its official split, with test-set captions as inputs and with no theatrical retrieval. TheaterAct is partitioned by play: the held-out play's plots, characters, and line sequences are unseen during training. Its captions are used for in-distribution theatrical evaluation; the remaining split is the sole retrieval pool, so no test-set motion is accessible at inference. Official HumanML3D checkpoints of MDM (Tevet et al., 2022) (475k), MLD (Chen et al., 2023), MoMask (Guo et al., 2024), and MMM (Pinyoanuntapong et al., 2024) are the caption-trained backbones. MDM is fine-tuned on TheaterAct for 1600 steps, 40% replay.

Metrics. For motion quality (Heusel et al., 2017; Guo et al., 2020; 2022), we use *FID* and *Diversity*. For physical plausibility (Yuan et al., 2023), we use *sliding*, *floating*, and *penetration*, all measured in world coordinates with the floor fixed at y=0. FID and related scores do not test whether acting intensity follows a requested dramatic axis, so **we propose two metrics: $\boldsymbol{\rho}$ and $\boldsymbol{R}_{\text{emo}}$.** A frozen encoder ϕ maps motion to a six-dimensional acting signal on $\mathcal{A}$; $\phi_{e^\star}$ is the score on the labeled axis:

$$\boldsymbol{\rho} = \text{Spearman}\big(\phi_{e^\star}(M_{\text{gen}}),\, \iota\big), \quad \boldsymbol{R}_{\text{emo}} = 100 \cdot \Pr_{M \sim \text{GT}^{e^\star}_{\setminus \text{test}}} \Big[\phi_{e^\star}(M) \le \phi_{e^\star}(M_{\text{gen}})\Big], \tag{9}$$

$\boldsymbol{\rho}$ checks that higher labeled intensity yields stronger acting on $e^\star$; $\boldsymbol{R}_{\text{emo}}$ places that strength within the empirical distribution of real performances on the same axis, so it is comparable across axes regardless of each axis's score scale. $\boldsymbol{R}_{\text{emo}}$=$50\%$ is the median real actor.

5.1 QUALITATIVE RESULTS

Figure 3 qualitatively evaluates script-driven acting generation on a held-out line, character, and Disgust intensity. Figure 3(a) feeds the *script* line to vanilla MDM, MLD, MoMask, and MMM, trained

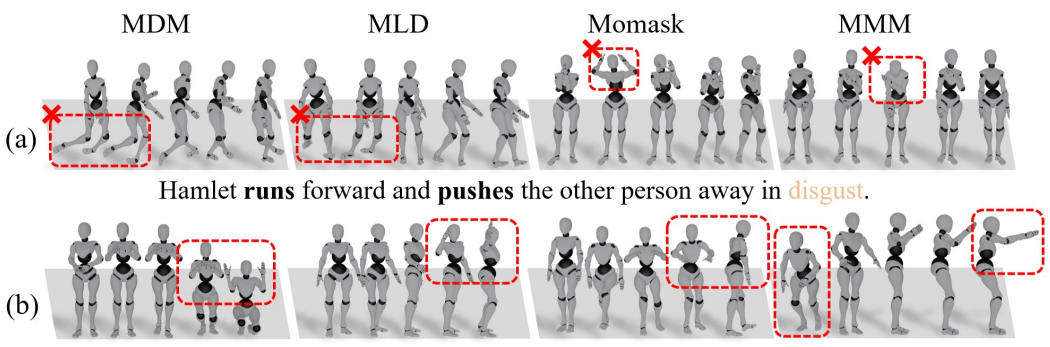

Figure 3: Held-out Hamlet line on Disgust. (a): caption-trained backbones queried with the script line; red marks indicate sliding or failed foot contact. (b): Script-to-Acting on MDM as τ increases, traversing distinct acting motifs on the same line.

on motion captions. This is a naive baseline for the script interface; their original caption setting is Table 2. The diffusion backbones fail on physics: vanilla MDM and MLD drag the supporting foot and float above the floor, with visibly deformed feet and lost ground contact over long spans, which matches the poor physical scores in Table 2. The masked-VQ backbones fail on content instead: MoMask and MMM keep the feet stably planted, but hold a single speaking pose for the entire line and never change it, so no dramatic beat is expressed. Both families therefore generate a single, content-consistent motion, and neither can play the line. Figure 3(b) is Script-to-Acting on the same MDM backbone and the same line: as τ increases, acting moves from kneeling prayer through one-hand rebuke and shove and turn to charging forward and shaking the other's shoulders. These are distinct motifs rather than one gesture rescaled, so τ selects how the line is played, not merely how strongly a fixed action is performed. Sweeping τ continuously, the same line yields a family of acting modes, and the foot-contact failures of the vanilla baselines no longer appear.

5.2 Quantitative Results

Stage acting serves as ground truth for testing script and emotion alignment; everyday-motion results on HumanML3D are in Appendix A.2. Table 2 compares Vanilla and +Ours on four backbones, diffusion (MDM, MLD) and masked VQ (MoMask, MMM). +Ours plugs in one fixed Acting Mind, Acting Morph, and Acting Creation stack without changing the backbone.

Acting Mind. Acting Mind encodes the script into a plan $\mathcal{T}$ rather than a motion caption, and $\mathcal{T}$ together with $(e^\star, \tau)$ enters the same Morph and Creation stack (Table 1). Supplying $\mathcal{T}$ raises all three acting metrics for every extractor. We report the extractor marked with * in Table 1, which is the one adopted for the final system and for all later evaluation: ρ from 0.61 to 0.63, R_{emo} from 67% to 76%, and Switch from 54.5% to 72.7%. The rises in R_{emo} and Switch quantitatively confirm our two central claims: emotion is expressed on the requested emotion axis, and sweeping τ moves the same line across distinct acting motifs rather than rescaling a single gesture. Figure 4 shows both on one line: in (a), where intensity is controlled by text alone, every generated sample stays inside a single motif cluster, whereas in (b), writing τ into retrieval lets the trajectory jump across clusters.

Acting Morph. Realizing $\mathcal{T}$ as a retrieved prefix makes that interface drop-in on all four backbones in Table 2, pulling generation toward the theatrical distribution. On MDM (5.47 to 6.35) and MMM (6.91 to 7.42), Diversity moves toward the GT value of 7.39. The physical metrics stay usable instead of degrading: sliding and penetration are severe for the Vanilla diffusion backbones and fall markedly with +Ours, while the masked-VQ backbones are already stable and no worse than Vanilla. Plug-and-play must therefore look like theater *and* improve the physical metrics.

Acting Creation. The frozen R_{emo} aligns emotion intensity on the requested axis: ρ is the rank correlation with τ, and R_{emo} measures where a sample falls within the same-axis distribution of real performances, where 50% is the median real actor and the GT row itself is 51%. On MDM, +Ours raises ρ from 0.52 to 0.63 and R_{emo} from 57% to $\mathbf{81}\%$, placing generated acting on that axis above four fifths of real performances. The other three backbones improve similarly, with R_{emo} reaching 72%, 71%, and 79%, and every backbone improves in ρ, most on MMM (from 0.50 to

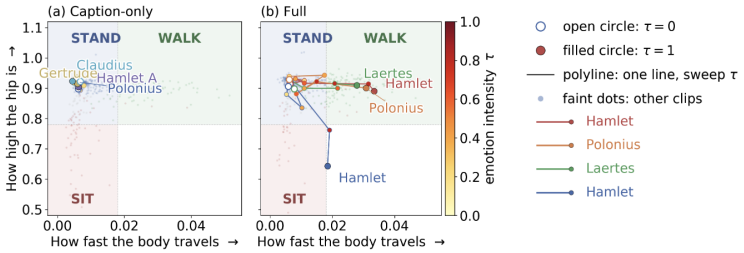

Figure 4: Sweeping τ on one line. (a) Text-only intensity stays inside one motion-motif cluster. (b) Writing τ into retrieval jumps the trajectory across clusters.

Table 1: Mind extractor on the TheaterAct test set (MDM backbone). Each metric compares *w/o* Mind (caption-only) with + Mind; * marks the extractor in the final system. Switch is the fraction of lines whose generated class changes when sweeping $\tau \in \{0, 0.2, \ldots, 1\}$. **Bold** / underline: best / second-best on the + Mind side, excluding FID.

	FID		Div ↑		ρ ↑		R_{emo} ↑		Switch ↑	
Mind	w/o	+	w/o	+	w/o	+	w/o	+	w/o	+
Qwen2.5-7B-Instruct*	16.09	7.26	4.75	**6.35**	0.61	**0.63**	67%	76%	54.5%	72.7%
GPT-5.6	16.15	6.70	4.34	5.92	0.48	0.52	69%	**81%**	50.6%	**75.3%**
DeepSeek-V3	16.31	7.25	4.51	6.31	0.56	0.62	70%	76%	50.6%	72.6%

0.68). Creation thus lets every backbone surpass the human sample distribution while sharpening the ordering along the emotion axis.

Table 2: Quantitative evaluation on TheaterAct. For each backbone, Vanilla and +Ours are shown side by side. **Bold** / underline: best / second-best excluding GT. FID is listed without ranking.

	FID		Diversity ↑		Sliding ↓		Floating ↓		Penetration ↓		ρ ↑		R_{emo} ↑	
Methods	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours
GT	—	—	7.39	—	5.0	—	3.20%	—	0.00%	—	0.61	—	51%	—
MDM	10.75	7.26	5.47	6.35	24.3	9.9	0.90%	3.84%	52.90%	9.82%	0.52	0.63	57%	**81%**
MoMask	6.17	5.44	7.02	6.64	6.5	6.3	1.50%	1.20%	5.50%	3.20%	0.45	0.53	46%	72%
MLD	7.39	5.55	6.57	6.65	23.5	12.8	**0.40%**	2.40%	45.60%	4.70%	0.57	0.61	45%	71%
MMM	6.25	1.95	6.91	**7.42**	6.7	**5.5**	1.20%	1.70%	**0.20%**	0.80%	0.50	**0.68**	43%	79%

5.3 Ablation Study

Acting Mind. Without $\mathcal{T}$, all three LLMs stay near FID 16, $R_{\text{emo}} \approx 67$–$70\%$, and Switch 50–55%; with $\mathcal{T}$ they reach FID ≈ 7, R_{emo} 76–81%, and Switch 72–75%. Figure 4 shows that caption-only results stay inside a single motion cluster, while adding Acting Mind leaves it. Swapping in GPT-5.6 and DeepSeek-V3 keeps the bands similar, so the lift comes from the structured plan rather than a stronger LLM; later modules read $\mathcal{T}$ rather than a particular extractor. As the table note states, FID is reported without ranking, and we make no higher-or-lower claim for it.

Acting Morph. *StyleOnly* broadcasts one global emotion code: it rescales a fixed motif but cannot switch it, giving the lowest diversity and the worst ground contact, while its low floating reflects a performance that barely moves. Its R_{emo} of 78% still approaches our 81%, since enlarging one motif already raises the encoder's score; ρ exposes this at 0.58, below the GT 0.61, so R_{emo} alone cannot certify content-variant control. *SingleMorph* adds the emotion and motif condition under a whole-line binding: professional priors alone already pull generation into the theatrical distribution and repair most ground contact. Its remaining gap is token-wise allocation, since it shares one condition across all tokens whereas we bind each separately.

Acting Creation. *NoEmoLoss* removes the frozen R_{emo}, leaving reconstruction and the retrieved prefix to fit alone: it leads on diversity and sliding, but floating becomes worst and ρ drops to 0.56, so motion still changes but no longer in order. Its R_{emo} of 46% matches the GT 51% against our 81%, showing the metric is governed by the emotion loss rather than the architecture, while our model is best on ρ (0.63), R_{emo} (81%), and penetration. The rows form a progressive verification chain rather than a set of independent baselines: motif priors, then token-wise allocation, then removal of the emotion loss. Content-variant, axis-ordered acting requires all three.

Table 3: Ablation on the TheaterAct test set, MDM backbone. Cond.: Global = a global style code, E+M = emotion and motif conditioning. Alloc.: Bcst. = the global code is broadcast to all tokens, Single = one motif bound to the whole line, Tok. = token-wise. **Bold** / underline: best / second-best excluding GT. FID is listed without ranking.

Methods	Module	Cond.	Alloc.	R_{emo}	FID	Div ↑	Slid. ↓	Flo. ↓	Pen. ↓	ρ ↑	R_{emo} ↑
GT	—	—	—	—	—	7.39	5.0	3.20%	0.00%	0.61	51%
StyleOnly	Morph	Global	Bcst.	✓	15.53	4.82	12.4	**2.50%**	14.80%	0.58	78%
SingleMorph	Morph	E+M	Single	✓	8.60	5.90	11.6	5.20%	11.20%	0.60	80%
NoEmoLoss	Creation	E+M	Tok.	×	6.15	**6.69**	**8.3**	9.17%	12.37%	0.56	46%
Script-to-Acting	All	E+M	Tok.	✓	7.26	6.35	9.9	3.84%	**9.82%**	**0.63**	**81%**

5.4 Human Evaluation

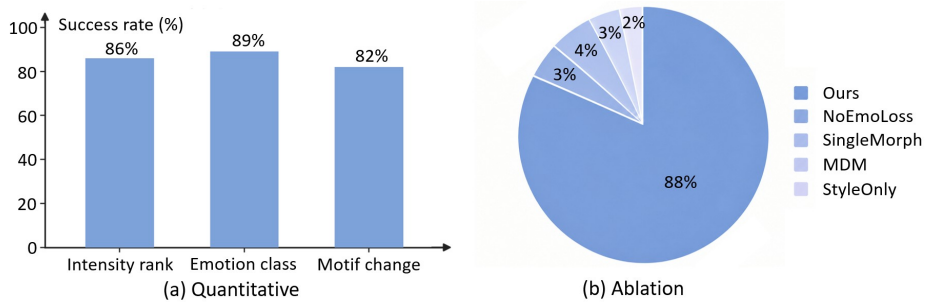

Figure 5: Human evaluation with 52 participants. (a) Success rate on three perceptual judgments. (b) Method win rate in four-way forced choices over the ablated variants.

We invited 52 participants to a blind study with hidden method names and randomly ordered clips. Figure 5(a) reports success rates on three judgments: the correct intensity ordering in 86% of trials, the intended emotion class in 89%, and a genuine change of acting motif in 82%. Together they show that the swept strength reads as continuous and ordered, the requested emotion is recognized on its own axis, and the change along τ is perceived as a different motif rather than a larger version of the same gesture. Figure 5(b) reports the overall preference rate, where our method is chosen in 88% of responses, against 4% for **SingleMorph**, 3% for **NoEmoLoss**, 3% for the text-to-motion paradigm, and 2% for **StyleOnly**. This ordering agrees with Table 3, confirming that the automatic metrics and human judgment are consistent.

6 Conclusion and Discussion

Conclusion. We presented Script-to-Acting, a unified framework for generating expressive, contextually aligned acting directly from scripts. Acting Mind, Acting Morph, and Acting Creation jointly enable continuous, content-variant control: changing emotion intensity can transform the action motif rather than merely rescaling a fixed motion. The modular design integrates with diverse motion-generation backbones and provides high-level, context-sensitive control over emotional and dramatic attributes. We further introduced TheaterAct, a multimodal dataset with rich annotations, establishing a foundation for script-conditioned acting generation.

Discussion. Our results suggest that emotion intensity in acting is better understood as a selection over a family of acting modes than as a global style gain: the same line can move across distinct motifs instead of only rescaling one motion. This points to a broader view of script-conditioned generation, in which the script is not merely a caption to be matched but a specification of dramatic intent that shapes which performance is produced. We hope this encourages work that treats affective and narrative structure as first-class control signals, and that connects language understanding with motion synthesis rather than applying one on top of the other. Our study further indicates that a small, carefully annotated theatrical corpus can expose behaviours that everyday-motion benchmarks leave invisible, which motivates domain-specific data as a driver of expressive generation.

AI USAGE STATEMENT

Large language models (LLMs) and multimodal large language models (MLLMs) are used as components of the methods evaluated in this work, including prompt rewriting, iterative refinement, and automatic evaluation, as described in the main text and appendix.

Generative AI tools were also used to assist with research implementation and manuscript preparation. In particular, AI-assisted tools were used to help write and debug scripts for video generation, evaluation, and iterative refinement, and to support grammar polishing, wording revisions, LaTeX formatting, and consistency checking. All experimental procedures, results, analyses, and final manuscript content were reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

We will release the code, model weights, sampled data, and evaluation scripts.

REFERENCES

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired motion style transfer from video to animation. *ACM Transactions On Graphics (TOG)* 39, 4 (2020), 64–1.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18000–18010.

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

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5142–5151.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM international conference on multimedia*. 2021–2029.

Ziyan Guo, Zeyu Hu, De Wen Soh, and Na Zhao. 2025a. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13869–13879.

Ziyu Guo, Young Yoon Lee, Joseph Liu, Yizhak Ben-Shabat, Victor Zordan, and Mubbasir Kapadia. 2025b. StyleMotif: Multi-Modal Motion Stylization using Style-Content Cross Fusion. *arXiv preprint arXiv:2503.21775* (2025).

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. 2023. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16750–16761.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems* 36 (2023), 20067–20079.

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

Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. 2023. Multiact: Long-term 3d human motion generation from multiple action labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1231–1239.

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

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems* 36 (2023), 25268–25280.

Qiujing Lu, Yipeng Zhang, Mingjian Lu, and Vwani Roychowdhury. 2022. Action-conditioned on-demand motion generation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2249–2257.

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

Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT Motion-Language Dataset. *Big Data* 4, 4 (dec 2016), 236–252. doi:10.1089/big.2016.0028

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. 2021. BABEL: Bodies, action and behavior with english labels. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 722–731.

Sigal Raab, Inbar Gat, Nathan Sala, Guy Tevet, Rotem Shalev-Arkushin, Ohad Fried, Amit Haim Bermano, and Daniel Cohen-Or. 2024. Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer. In *SIGGRAPH Asia 2024 Conference Papers*. 1–13.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. 2024. World-Grounded Human Motion Recovery via Gravity-View Coordinates. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.

Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. 2024. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 821–830.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 448–458.

Qing Wang, Xiaohang Yang, Yilan Dong, Naveen Raj Govindaraj, Gregory Slabaugh, and Shanxin Yuan. 2025. Dancechat: Large language model-guided music-to-dance generation. *arXiv preprint arXiv:2506.10574* (2025).

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. 2025. Motionagent: A conversational framework for human motion generation with llms. In *International Conference on Learning Representations*, Vol. 2025. 48023–48043.

Tan Yu, Jingjing Wang, Jiawen Wang, Jiamin Luo, and Guodong Zhou. 2024. Towards emotion-enriched text-to-motion generation via LLM-guided limb-level emotion manipulating. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 612–621.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2023. Physdiff: Physics-guided human motion diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 15964–15975.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence* 46, 6 (2024), 4115–4128.

Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2024. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*. Springer, 405–421.

Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. 2024. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*. Springer, 18–38.

Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10544–10553.

A Appendix

A.1 Additional Qualitative Results

Figure 6 extends the qualitative comparison of Section 5.1 to all six emotion axes. Each row sets one active axis and shows the generated performance for the corresponding held-out line, together with the motion caption of its realized acting. The realized acting differs across the six rows, which is the per-axis behaviour the main text reports for Disgust alone.

A.2 Quantitative Experiments on HumanML3D

We evaluate the same acting-aligned checkpoints on the HumanML3D test set, without theatrical retrieval, as an out-of-domain check against everyday motion rather than as the headline result. The generator is aligned to professional acting; HumanML3D ground truth is everyday action. Table 4 uses the same Vanilla / +Ours layout as Table 2. ρ and R_{emo} are omitted: HumanML3D has no acting axis and no intensity τ. FID is listed without ranking. The GT FID cell is left empty because FID is a distance to the real distribution; GT compared with itself is identically zero.

The headline on this split is the expected domain shift, not a failure of the main table. Vanilla stays closer to the HumanML3D distribution; theatrical fine-tuning pulls generation away from short everyday clips. Script alignment and emotion alignment are still decided on TheaterAct.

MDM shows the largest shift. The official HumanML3D checkpoint reports FID 0.54 and diversity 9.56, matching GT diversity 9.51. After +Ours, FID is 13.12 and diversity 5.81: acting fine-tuning turns the diffusion prior toward performance and forgets part of the everyday repertoire. Physical metrics improve at the same time (sliding $33.9 \rightarrow 12.7$, floating $8.2\% \rightarrow 5.2\%$, penetration $26.6\% \rightarrow 13.7\%$). On HumanML3D, MDM +Ours is less like everyday motion and less prone to skating and penetration, which is consistent with acting alignment.

MoMask barely leaves the everyday prior. Vanilla FID is 0.12 and +Ours is 0.081; diversity moves from 9.36 to 9.54; sliding / floating / penetration are essentially unchanged (8.8 / 10.4% / 1.4% vs. 8.4 / 10.0% / 1.5%). A short fine-tune is enough to inject the acting signal and not enough to overwrite HumanML3D. MoMask is the most stable backbone on this check.

MLD sits in between. FID rises from 6.21 to 9.18 and diversity dips from 7.97 to 7.23, but sliding falls from 30.5 to 18.7 and penetration from 39.6% to 17.0%. Under the readable physics protocol, official MLD already skates and penetrates heavily; acting fine-tuning does not restore everyday FID, yet it tightens contact with the ground.

MMM is similar. Vanilla FID 0.11 and diversity 9.65 already sit near GT; +Ours raises FID to 2.65 while diversity stays high (9.39). Sliding improves $8.6 \rightarrow 6.1$ and floating $7.7\% \rightarrow 6.0\%$; penetration is 0.2% vs. 0.3%. MMM +Ours is strong on the theatrical main table; on everyday motion it pays a modest distribution shift without worse physics.

Across the four backbones, +Ours is not a method for HumanML3D FID. The more sensitive the everyday prior (MDM), the larger the FID move; the more stable the discrete tokenizer and the

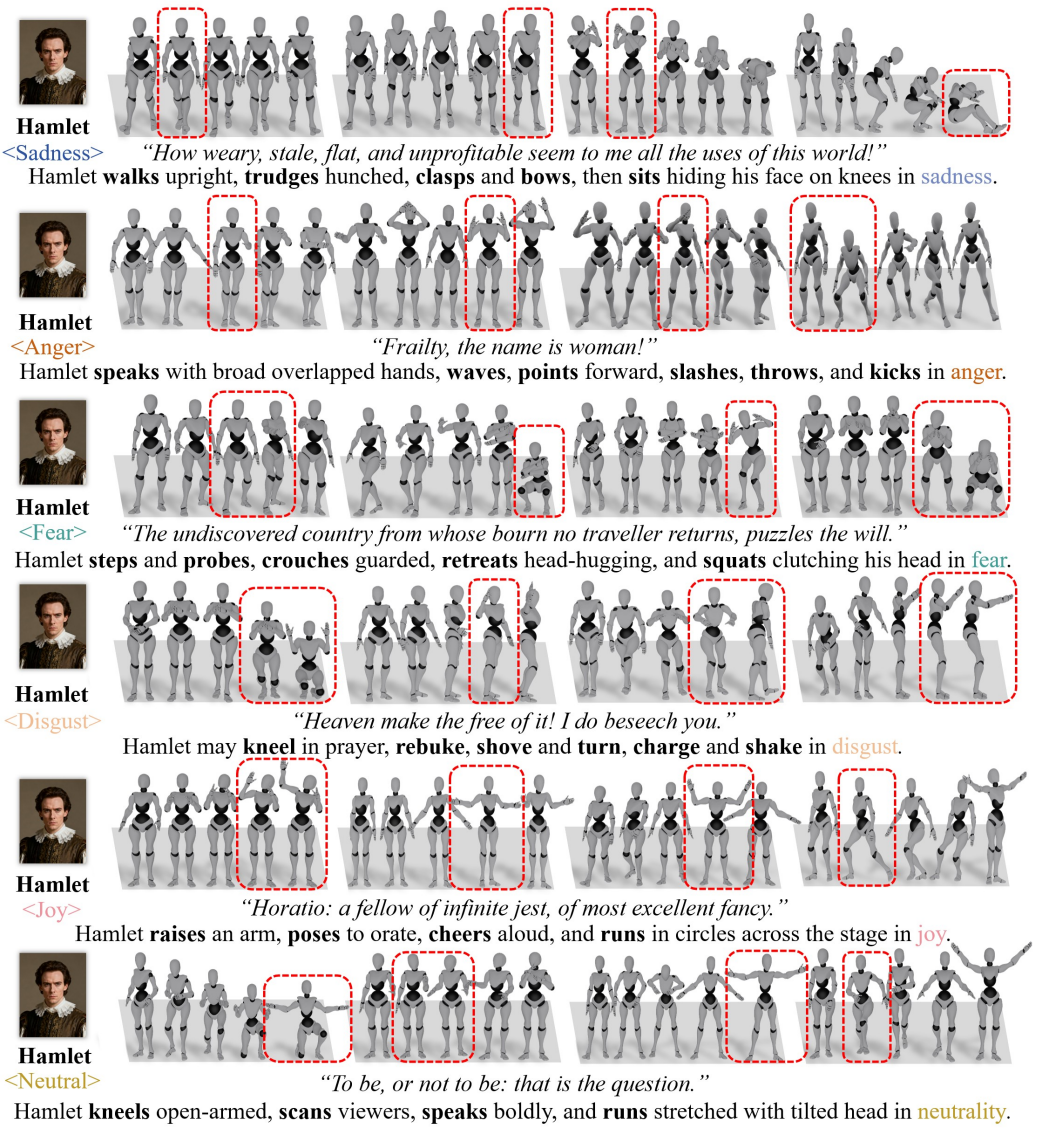

Figure 6: Qualitative results on held-out Hamlet material, one row per emotion axis. Each row shows sampled frames of the generated performance, with the action cues of its realized acting marked in bold. The same character and the same script excerpt yield a distinct acting motif on every axis: hidden-face grief under Sadness, frantic waving, throwing, and kicking under Anger, a guarded head-clutching crouch under Fear, shove-away, turn, and confrontation under Disgust, raised arms and circling jumps under Joy, and open-armed scanning under Neutral. The Disgust row uses the same held-out line as Figure 3.

shorter the fine-tune (MoMask), the less everyday metrics move. On physics, +Ours is no worse than Vanilla and often better: acting alignment does not turn motion into heavier skating or penetration. HumanML3D only shows that the fine-tuned generator still produces evaluable everyday clips; acting quality is reported on TheaterAct.

Table 4: Evaluation on HumanML3D test. Same V3 weights as TheaterAct, no theatrical retrieval. For each backbone, Vanilla and +Ours are shown side by side. **Bold** / underline: best / second-best excluding GT. FID is listed without ranking. ρ and R_{emo} do not apply (no emotion labels).

	FID		Diversity ↑		Sliding ↓		Floating ↓		Penetration ↓		ρ ↑		R_{emo}	
Methods	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours	Vanilla	+Ours
GT	—	—	9.51	—	6.7	—	8.7%	—	0.00%	—	—	—	—	—
MDM	0.54	13.12	9.56	5.81	33.9	12.7	8.2%	5.2%	26.6%	13.7%	—	—	—	—
MoMask	0.12	0.081	9.36	9.54	8.8	8.4	10.4%	10.0%	1.4%	1.5%	—	—	—	—
MLD	6.21	9.18	7.97	7.23	30.5	18.7	**5.1%**	5.4%	39.6%	17.0%	—	—	—	—
MMM	0.11	2.65	**9.65**	9.39	8.6	**6.1**	7.7%	6.0%	**0.2%**	0.3%	—	—	—	—

A.3 THEATERACT DATASET

A.3.1 SCALE AND DURATION CHARACTERISTICS

As a supplement to Section 4, we detail the composition of the *TheaterAct* dataset. The release comprises professional acting sequences totalling **7.2 hours** at 20 fps (730 clips; mean length 35.5 s), cropped from about **30 hours** of raw theatrical video and aligned to the HumanML3D 263-dimensional joint representation. It covers **10** plays and **59** character cards with distinct occupations, personalities, and motivations. Sequence length is highly concentrated: 545 of the 730 clips last 35–40 s, and with 91 clips at 25–30 s and 83 at 30–35 s, almost every clip falls between 25 s and 40 s, while only 9 are shorter than 25 s, as shown in Figure 7(b). Compared with mundane motion–language corpora whose clips last 2–10 s, *TheaterAct* gives a line room to evolve on one emotion axis rather than collapsing into a single atomic motif.

A.3.2 COMPARISON WITH EXISTING DATASETS

As shown in Table 5, *TheaterAct* differs from existing datasets in two respects. First, HumanML3D (Guo et al., 2022), KIT-ML (Plappert et al., 2016), Motion-X (Lin et al., 2023), and BABEL (Punnakkal et al., 2021) are sourced from everyday MoCap or in-the-wild video, so motion style is decoupled from dramatic delivery. *TheaterAct* takes theatrical video as its source, so both the motion and the emotional delivery are those of staged acting. Second, those corpora annotate a text–motion pair (a caption or an action tag) and do not provide the theatrical knowledge required for acting generation: a play script, spoken lines, or scene context, nor acting emotion, intensity, or motivation. *TheaterAct* aligns each clip with a line of dialogue, a HumanML3D-style motion caption, and the enclosing play script (including character identity C). Expert labels are six emotion types, a scalar intensity τ on the active axis, and ten character motivations. Character identity is carried by the script, not listed as a separate expert head. This is the data substrate for *script alignment* and *emotion alignment* in the main paper. Because the release is deliberately concentrated on staged performance, its 7.2 hours are not directly comparable in size to the 11–144 hours of the general-purpose corpora; the gain is in annotation depth rather than raw duration.

A.3.3 GENRE DIVERSITY AND SOCIAL IDENTITY

TheaterAct is built to capture the range of expressiveness in theatrical performance. Clips are retained from the source video after cropping and 20 fps alignment, so the released 7.2 hours concentrates staged acting rather than uncut production footage.

Each of the 10 plays is labelled with a drama category in its script-level annotation; across the release these cover Tragedy, Comedy, and Tragicomedy, and the schema additionally distinguishes Historical Drama, Romance, Satire, Absurd Drama, and Realist Drama. Character background is recorded through occupation: of the 59 character cards, 25 are Commoner, 16 Noble, 8 Worker, 5 Artist, 3 Soldier, and 2 Merchant, as shown in Figure 7(a). This annotation lets the model condition on given circumstances—for example the bearing of a noble versus that of a commoner—and synthesize character-driven acting rather than a generic trajectory. Motivation is annotated as ten tags (Pursuing Love, Altruistic, Escaping, Loyal, Moral, Pursuing Power, Manipulative, Selfish, Self-Sacrifice, Seeking Revenge), whose distribution is reported in Figure 8(b). Motivation is a multi-label *character* attribute, not a per-line tag: the shares are computed over tag occurrences, the most frequent tag accounts for 18% of occurrences and the rarest for 3%, so no single motivation

dominates the corpus. Together with personality keywords and a short description, motivation is part of the given circumstances of C that ActingMind and retrieval read as script context.

A.3.4 Dataset Annotation and Organization

To support script-driven theatrical motion generation, annotations are organized at the play, character, and line levels. Each play has one `script_english.json` with three blocks: `play`, `characters`, and `lines`.

```
script_english.json
|-- play
|   |-- name, time, plot_summary, props, drama_category
|   `-- see the script-level schema table
|-- characters
|   |-- character
|   |   |-- name, description, personality, motivation
|   |   `-- occupation, social_identity, economic_status, education
|   `-- see the character-level schema table
`-- lines
    |-- character_name
    |   `-- linesID, lines, text
    `-- see the dialogue-level schema table
```

The schemas are summarized in Tables 6–8.

The `play` block stores the title, period, plot summary, key props, and drama category. These fields give the global circumstances of the scene: setting, period, and play type.

The `characters` block stores, for each role, a short description, personality keywords, occupation, and a subset of the ten motivations. These fields are the given circumstances of character C. Cards also record gender, economic status, and education; this paper uses occupation as the social attribute (Figure 7(a)).

The `lines` block is the clip-level unit. Each entry links a dialogue span (with timing marks) to a 263-D motion file and a HumanML3D-style caption. A sidecar emotion file adds the active axis $e^\star$ among $\mathcal{A} = \{\text{Anger}, \text{Disgust}, \text{Fear}, \text{Joy}, \text{Sadness}, \text{Neutral}\}$ and an intensity $\tau \in [0, 1]$ written onto that axis (zeros elsewhere). Rare fine tags Surprise and Contempt, when present, are merged into Joy and Disgust respectively. The dialogue-level schema also records eight dramatic emotion axes with continuous intensities, but we do *not* use this eight-pair emotion vector as a generation condition in this paper, and the release does not include audio.

The directory layout is as follows.

```
TheaterAct/
|-- play/
|   |-- script_english.json
|   |-- character/
|   |   |-- emotion/
|   |   |-- line_en/
|   |   |-- new_joint_vecs/
|   |   `-- text/
|   `-- ...
`-- ...
```

Each of the 10 plays stores `script_english.json` and one folder per character. Under `play/character/` sit dialogue transcripts (`line_en`), motion captions (`text`), emotion sidecars (`emotion`), and `new_joint_vecs` (263-D, 20 fps). That correspondence among script, character, line, and motion is what ActingMind reads as $\mathcal{T} + (e^\star, \tau)$ and what ActingMorph indexes (training plays only).

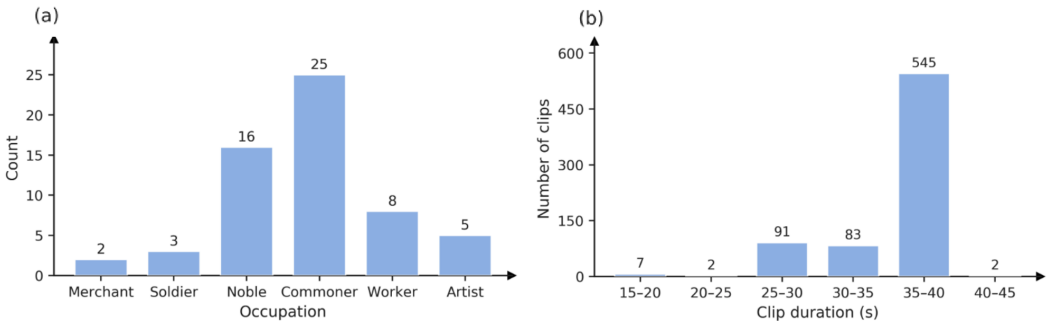

Figure 7: (a) Occupation of the 59 character cards. (b) Clip duration at 20 fps (730 clips; mean 35.5 s), mostly 35–40 s.

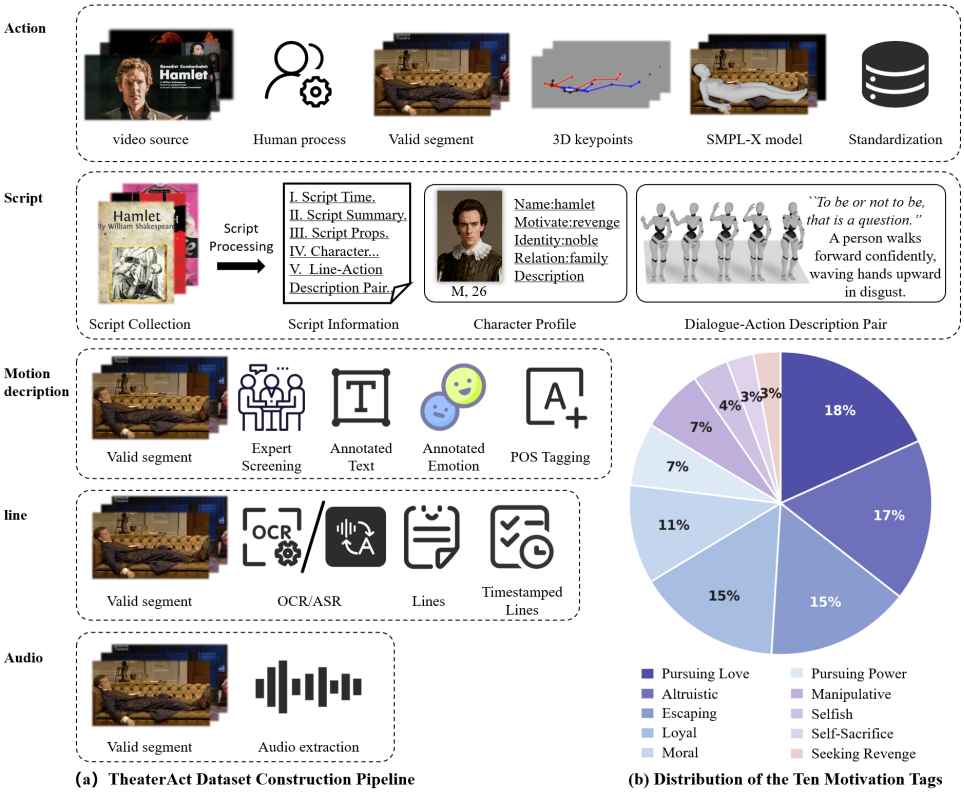

Figure 8: TheaterAct construction and annotation statistics. (a) The dataset construction pipeline, from theatrical source video to standardized 263-dimensional motion. (b) Distribution of the ten motivation tags across the 59 character cards. Motivation is a multi-label character attribute, not a per-line tag, so the shares are computed over tag occurrences.

A.3.5 Partitioning Strategy for Generalization

We partition *TheaterAct* by play rather than by a random clip split. The test play's plot, character relations, and line order are entirely unseen during training, and never enter the retrieval index or checkpoints. The model must read a new script rather than replay memorized sequences.

Table 5: **Dataset comparison.** *TheaterAct* provides theatrical motions with aligned scripts and expert acting labels. Character identity is carried by the script, not a separate expert label.

				Script information			Expert annotations		
Dataset	**Source**	**Duration**	**Text**	**Script**	**Dialogue**	**Scene**	**Emotion**	**Intensity**	**Motivation**
HumanML3D (Guo et al., 2022)	MoCap	28.6h	✓	×	×	×	×	×	×
KIT-ML (Plappert et al., 2016)	MoCap	11.2h	✓	×	×	×	×	×	×
Motion-X (Lin et al., 2023)	Video	144.2h	✓	×	×	×	×	×	×
BABEL (Punnakkal et al., 2021)	MoCap	43.5h	✓	×	×	×	×	×	×
TheaterAct (Ours)	**TheaterVid**	**7.2h**	✓	✓	✓	✓	**6 Types**	✓	**10 Kinds**

Table 6: Script-level annotation schema of the TheaterAct dataset.

Field	**Values / Description**	**Purpose**
Script ID	Unique identifier	Primary key
Script Name	Full script title	Script identification
Drama Category	Tragedy, Comedy, Historical Drama, Romance, Satire, Tragicomedy, Absurd Drama, Realist Drama	Defines dramatic genre
Drama Style	Classicism, Romanticism, Realism, Expressionism, Symbolism, Postmodernism	Defines artistic style
Plot Summary	∼300-word summary emphasizing the central emotion	Narrative overview
Core conflicts	Human vs. Society, Human vs. Human, Human vs. Self, Human vs. Fate, Human vs. Environment	Defines the primary dramatic emotion
Drama Theme	Love, Power, Revenge, Redemption, Growth, Death, Justice, Freedom	Defines the narrative theme
Time Background	Ancient, Medieval, Renaissance, Modern, Contemporary, Future	Temporal background
Props	Key props affecting character performance; *null* if unavailable	Performance context
Time Setting	Specific time setting; *null* if unavailable	Temporal context

Table 7: Character-level annotation schema of the TheaterAct dataset.

Field	**Values / Description**	**Purpose**
Character ID	Unique identifier	Primary key
Character Name	Full character name	Character identification
Social Identity	Upper Class, Middle Class, Lower Class, Student, Special Occupation, Unemployed	Defines social background
Gender	Male, Female, Elderly, Young	Character attribute
Character Type	Protagonist, Supporting Character, Antagonist, Comic Character, Tragic Character, Functional Character	Defines dramatic role
Personality Traits	3–5 descriptive keywords (e.g., Brave, Intelligent, Hesitant, Loyal)	Personality representation
Background Story	∼200-word summary of key experiences	Behavioral motivation
Main Relationships	Relationships with other characters	Social interactions
Relationship Type	Kinship, Power, Friendship, Love, Mentor–Student, Hostility, Employment, Master–Servant, Colleague, Stranger	Defines relationship category
Relationship Dynamic	Stable, Improving, Deteriorating, Complex	Relationship evolution
Relationship Strength	Strong, Medium, Weak	Relationship intensity

Table 8: Dialogue-level annotation schema of the TheaterAct dataset.

Field	Values / Description	Purpose
Dialogue ID	Unique identifier	Primary key
Dramatic Emotions	Victory vs. Defeat, Calmness vs. Frenzy, Love vs. Hatred, Joy vs. Sorrow, Courage vs. Fear, Confidence vs. Cowardice, Hope vs. Despair, Anger vs. Composure. Each axis is annotated with a continuous intensity score ranging from 0.0 to 1.0 (step = 0.1).	Fine-grained emotion modeling
Expression Mode	Rational, Emotional, Paranoid, Suspicious, Innocent, Sophisticated, Reserved, Expressive	Expressive characterization
Motivation	Revenge, Love, Power, Wealth, Honor, Freedom, Survival, Redemption, Protection, Destruction	Behavioral intention
Dialogue	Original script dialogue (e.g., "To be, or not to be...")	Linguistic context
Scene	Scene description (e.g., castle corridor, battlefield, bedroom)	Environmental context
Script ID	Unique identifier linked to the Script Information table	Script association
Character ID	Unique identifier linked to the Character Information table	Character association
Text	HumanML3D-style motion description (e.g., "A person walks slowly...")	Motion-language compatibility
Motion File	263-D joint representation (.npy)	Motion representation
Interaction Type	Monologue, Dialogue	Interaction modeling

Table 9: Symbols used in Acting Mind, Acting Morph, and Acting Creation.

Symbol	Description
Inputs and emotion axes	
$\mathcal{T}_{\text{script}}$	Raw play script.
C	User-chosen speaker (character).
$e^\star$	Active emotion axis, $e^\star \in \mathcal{A}$.
τ	Requested intensity on $e^\star$, $\tau \in [0, 1]$.
$\mathcal{A}$	$\{\text{Anger}, \text{Disgust}, \text{Fear}, \text{Joy}, \text{Sadness}, \text{Neutral}\}$.
Acting Mind (§3.1)	
Φ_{LLM}	Dramaturgical context projector (an LLM agent).
Emo	Emotion vector $\text{Emo} \in [0, 1]^6$; only the active coordinate is overwritten, $\text{Emo}_{e^\star} = \tau$.
$\mathcal{T}$	Acting plan $\{C, \text{line}, e^\star, \text{Emo}\}$; it contains no demonstrated joint trajectory.
Acting Morph (§3.2)	
H_s	Script-token features of the line, $H_s \in \mathbb{R}^{L_s \times d}$.
$e_q^{\text{sem}}, e_i^{\text{sem}}$	Frozen CLIP text embeddings of the query line and of professional clip i.
q_n	Token-level query built from the n-th script token.
a_i, ι_i	Annotated active axis and intensity of clip i.
$\omega_{\text{sem}}, \omega_{\text{emo}}$	Weights on semantic cosine similarity and on on-axis intensity match.
$\text{Sim}(q, i)$	Retrieval score of query q against clip i; Eq. (3).
$\varepsilon_s, \varepsilon_p$	Thresholds on foot sliding and ground penetration; Eq. (4).
$\text{Fil}_{\text{phy}}(i)$	Contact-geometry filter on sliding and ground penetration; Eq. (4).
S	Token-wise acting-mode matrix $S_{n,k} = \text{Sim}(q_n, i_k)$: binding strength of token n to motif k; Eq. (5).
$\mathcal{M}_{\text{ref}}(\tau)$	Clip set kept as motif columns after Top-$K(S)$ and Fil_{phy}; M_{ref} denotes the retrieved motion itself.
H_{ref}	Encoded selected motif columns, $H_{\text{ref}} \in \mathbb{R}^{L_r \times d}$; the only Acting Morph output handed to Acting Creation.
Acting Creation (§3.3)	
H_m	Motion-token features of the backbone.
X	Self-attention input $[H_{\text{ref}}; H_m]$; realized by the backbone's existing layers.
m_{ref}	Mask of the retrieved prefix frames used as the anchor.
$x_0, x_t, \hat{x}_0$	Clean motion window, noisy input at diffusion step t, and denoiser prediction.
D_θ	Trainable motion generator; gradients update only D_θ.
$\mathcal{L}_{\text{rec}}$	Reconstruction term $\mathbb{E}\|\hat{x}_0 - x_0\|_2^2$; Eq. (6).
$\mathcal{L}_{\text{ref}}$	Retrieved-prior anchor term $\mathbb{E}\|(\hat{x}_0 - M_{\text{ref}}) \odot m_{\text{ref}}\|_2^2$; Eq. (6).
$\mathcal{L}_{\text{emo}}$	Emotion term $-\varphi_{e^\star}(\hat{x}_0)$ on the active axis; Eq. (7).
$\lambda_{\text{rec}}, \lambda_{\text{ref}}, \gamma$	Weights in $\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{ref}}\mathcal{L}_{\text{ref}} + \gamma\mathcal{L}_{\text{emo}}$; Eq. (8).
φ	Frozen motion-to-emotion encoder, reused by $\mathcal{L}_{\text{emo}}$, ρ, and R_{emo}.
Metrics (§5)	
ρ	Spearman rank correlation between $\varphi_{e^\star}(M_{\text{gen}})$ and ι.
R_{emo}	Percentile of $\varphi_{e^\star}(M_{\text{gen}})$ within real non-test performances on the same axis; 50 = median real actor.

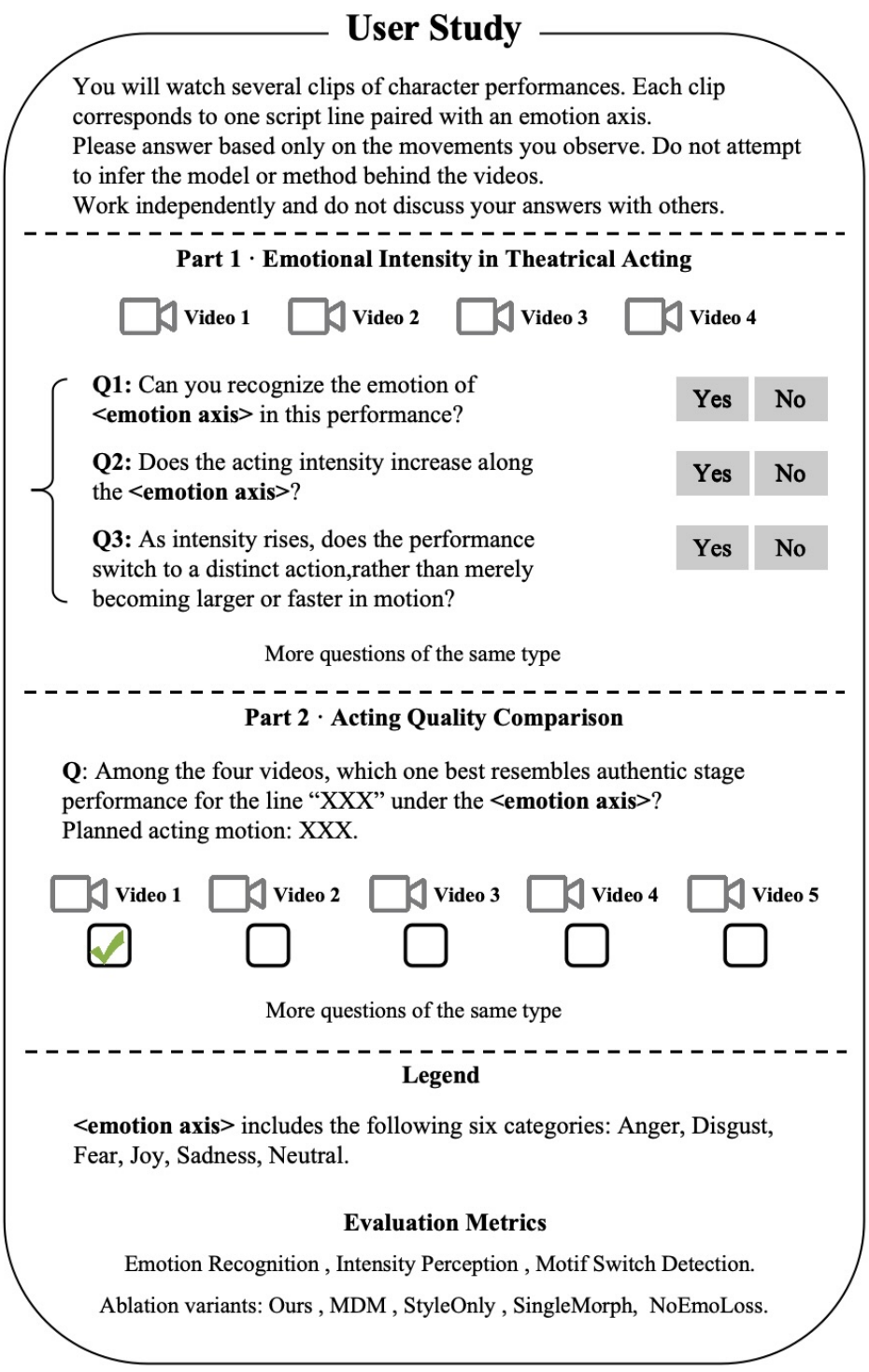

Figure 9: Human evaluation questionnaire. Method names are hidden and clips are presented in random order. **Part 1** shows four clips of the same line under one emotion axis and asks three binary judgments: whether the intended emotion is recognizable, whether the acting intensity increases along that axis, and whether the performance switches to a distinct action rather than merely becoming larger or faster. **Part 2** shows five clips of the same line and asks which one best resembles authentic stage performance.