EmoText-Talk: Text-Driven Emotion and Intensity Control for Talking Head Generation via Disentangled CLIP Routing

Anonymous ACL submission

Abstract

Existing emotional talking-head methods require reference videos or rigid category labels, limiting flexibility and demanding paired data at inference. We propose EmoText-Talk, a lightweight, 2.5M-parameter plug-in that adds free-form, text-driven control over both emotion and intensity to a frozen EDTalk generator, preserving its identity, lip-sync, and visual quality without retraining. Our key insight is to separate *what* emotion to express from *how strongly* via two novel components. First, since EDTalk's compact expression space hides emotion directions its reconstruction loss cannot recover, we discover per-emotion targets through a classifier-guided search over MEAD actors and regress CLIP text embeddings onto them. Second, an IntensityAwareProjection decorrelates intensity from emotion, dropping their CLIP cosine similarity from 0.976 to 0.046, so a continuous scalar predicted from language controls expression magnitude. Supporting CLIP-PAE projection and Dynamic Graph Attention routing over eight emotion nodes map each prompt to its category. EmoText-Talk reaches 91.67% emotion accuracy on MEAD, surpassing prior text-driven methods while keeping the generator frozen.

1 Introduction

Generating realistic talking head videos with controllable facial expressions has broad applications in virtual avatars, film production, video conferencing, and human-computer interaction (Tan et al., 2024; Gan et al., 2023; Zhang et al., 2023). Recent audio-driven methods (Zhang et al., 2023; Prajwal et al., 2020; Xu et al., 2024; Yang et al., 2025) achieve impressive lip synchronization and fidelity, but controlling the emotional expression of generated faces remains challenging.

Existing emotional methods (Gan et al., 2023; Ji et al., 2022; Tan et al., 2024; Ma et al., 2023b; Wang et al., 2023) condition expression on one of three signals. Categorical labels lack nuance, reference emotion videos demand paired footage at inference and cannot generalize, and audio-derived features are limited by the emotional content of the audio. However, they cannot support the flexible control that free-form text provides.

We propose EmoText-Talk, a novel method for zero-shot, *text-driven* emotion and intensity control. Given a prompt such as "*The person is expressing anger with high intensity*," it generates a video with the corresponding expression while preserving source identity and lip sync. Our approach is a lightweight plug-in ($\sim$2.5M parameters) on a frozen EDTalk (Tan et al., 2024) generator, preserving all base-model qualities (lip sync, identity, pose) while adding text-driven emotion control.

Delivering this control raises three requirements that off-the-shelf components do not meet. First, CLIP (Radford et al., 2021) embeddings for different emotions are nearly collinear (cosine 0.88 to 0.95) and barely encode intensity (0.976 between low and high intensity prompts), so we amplify the discriminative component with CLIP-PAE subspace projection (Zhou et al., 2023) and route it through a Dynamic Graph Attention network (Ding et al., 2025) over eight emotion nodes. Second, EDTalk's 10-D expression latent hides emotion along directions its reconstruction loss never recovers. Our first novelty, a *classifier-guided search*, discovers these targets across 18 MEAD actors and regresses text embeddings onto them, the single most impactful component of the system. Third, emotion and intensity share a single prompt yet must be controlled independently. Our second novelty, an *IntensityAwareProjection*, decorrelates the two (CLIP cosine 0.976 to 0.046) and predicts a continuous intensity scalar from natural language alone.

Our contributions are twofold–

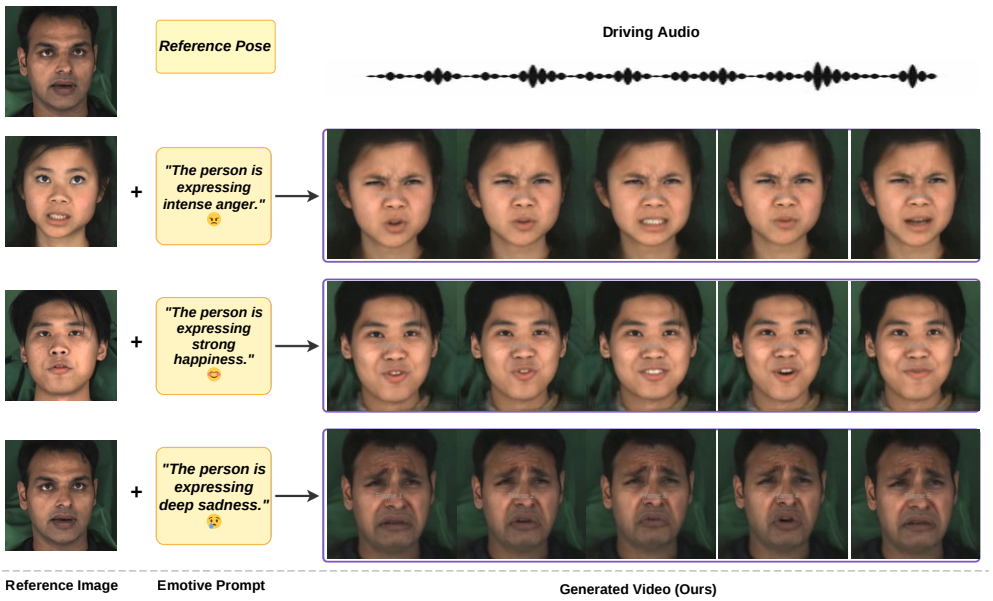

Figure 1: Text-driven emotion control with EmoText-Talk. Given a reference identity image, a reference pose guide, a driving audio signal, and a natural-language emotive prompt (*e.g.,* "the person is expressing intense anger," "the man exhibits strong happiness"), our method generates a talking-head video whose facial expression matches the described emotion while preserving identity and synchronising lip motion with the audio.

- EmoText-Talk, the first text-driven emotion-and-intensity control module for talking head generation, giving zero-shot control without reference videos or retraining, built on classifier-guided expression discovery and an intensity-disentangling projection.
- State-of-the-art emotional talking-head generation with explicit, fine-grained control of emotion intensity on a frozen generator.

On MEAD, EmoText-Talk reaches 91.67% accuracy (+22.8 over EDTalk-V, above EmoCAST's 83.60%) with lip-sync and identity unchanged. Ablations confirm both novelties. Expression discovery is the largest single factor (over 60 points, matching CLIP-PAE) and the IntensityAwareProjection gives correct ordinal intensity for all seven emotions, while attention-supervised DGAT routing adds 14.6 points (8/8 correct).

2 Related Work

Audio-Driven Talking Head Generation. Methods have progressed from lip-sync approaches (Prajwal et al., 2020) to full face animation. SadTalker (Zhang et al., 2023) predicts 3D Morphable Model (3DMM) coefficients from audio. KDTalker (Yang et al., 2025) combines implicit 3D keypoints with spatiotemporal diffusion for pose diversity. Diffusion-based methods like Hallo (Xu et al., 2024), EMO (Tian et al., 2024), MEMO (Zheng et al., 2024) achieve high fidelity at heavy cost. EDTalk (Tan et al., 2024) disentangles lip, pose, and expression into orthogonal latent spaces for independent control. We build on EDTalk, adding text-driven emotion control to its expression pathway while keeping the generator frozen.

Emotion-Aware Talking Head Generation. EAMM (Ji et al., 2022) transfers emotion from an emotion source video as additive motion displacements. EAT (Gan et al., 2023) maps labels and intensity levels to latent codes. PD-FGC (Wang et al., 2023) uses progressive disentanglement with reference videos, and StyleTalk (Ma et al., 2023b) transfers styles via style codes. EDTalk (Tan et al., 2024) uses predefined per-emotion weights (68.85% Acc_{emo}, the best video-driven result at the time). GMTalker (Xia et al., 2024) uses Gaussian mixture variational autoencoders (VAEs), EMOdiffhead (Zhang et al., 2024) enables continuous diffusion control, and EmoHead (Shen et al., 2025) probes hyperplanes in FLAME space (75.98%). All rely on reference videos or labels at inference, whereas we instead use natural language at higher accuracy.

Text-Driven Facial Expression Control. TalkCLIP (Ma et al., 2023a) trains a CLIP style encoder but targets speaking style, not discrete emotion.

ExpCLIP (Zhong et al., 2024) aligns text and expressions on TEAD, but is unreleased. EmoCAST (Jiang et al., 2025) uses a diffusion framework with text-guided emotive attention (83.60%, prior text-driven state of the art) but trains a full Stable Diffusion pipeline from scratch on a custom dataset (Hallo3 + CelebV-HQ). We instead map text to an existing generator's expression space with a ∼2.5M-parameter plug-in, preserving pretrained quality at higher accuracy.

Limitations of Current Research. No prior approach offers intensity-aware emotion control from free-form text. Audio-driven methods do not control emotion. Emotion-aware methods rely on coarse labels or reference videos that need paired footage and leak identity. Text-driven methods so far handle only speaking style, stay unreleased, or train a costly diffusion model from scratch. EmoText-Talk closes this gap, mapping free-form text onto a frozen generator's expression space while keeping emotion and intensity separate.

3 Method

3.1 Overview

Given a source image I^s, audio A, pose video P, and a prompt T (*e.g.*, "*anger with high intensity*"), we generate a video where the face expresses the described emotion at the right intensity while keeping lip sync and identity. As shown in Figure 2, EmoText-Talk maps prompts to 10-D expression weights via two disentangled pathways, the first an *emotion direction* pathway (*which* emotion) and the second an *intensity* pathway (*how strongly*), both of which are integrated with a frozen EDTalk generator.

$$\begin{aligned} \mathbf{w}_{\text{base}} &= \text{Proj}(\text{DisTIB}(\text{CrossAttn}(\text{PAE}(\mathbf{e}), \\ &\qquad \text{DGAT}(\mathbf{e})))) \\ s &= \text{IntHead}(\text{IntProj}(\mathbf{e})) \\ \mathbf{w}_{\text{final}} &= \mathbf{w}_{\text{base}} \cdot s \end{aligned} \tag{1}$$

where $\mathbf{e} = \text{CLIP}_{\text{text}}(T) \in \mathbb{R}^{512}$ is the frozen CLIP text embedding.

3.2 CLIP Text Encoding and Subspace Projection

We encode T with a frozen CLIP ViT-B/32 (Radford et al., 2021) text encoder to obtain $\mathbf{e} \in \mathbb{R}^{512}$. CLIP embeddings for different emotions are nearly identical (cosine 0.88 to 0.95). Prompts "a person expressing *[emotion]*" differ only in the emotion word, which CLIP's bag-of-words behavior (Koishigarina et al., 2026) suppresses. Following CLIP-PAE (Zhou et al., 2023), we compute 8 prototype embeddings $\{\mathbf{p}_k\}_{k=1}^{8}$ (angry, contempt, disgusted, fear, happy, neutral, sad, surprised) and extract an orthonormal basis $\mathbf{V}_8 \in \mathbb{R}^{8\times 512}$ via SVD.

$$\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top = \text{SVD}([\mathbf{p}_1, \ldots, \mathbf{p}_8]^\top), \quad \mathbf{V}_8 = \mathbf{V}_{:8}^\top \tag{2}$$

We decompose the input into an emotion projection and residual, then amplify the discriminative component.

$$\mathbf{e}_{\text{proj}} = \mathbf{V}_8^\top \mathbf{V}_8 \cdot \mathbf{e}, \quad \mathbf{e}_{\text{res}} = \mathbf{e} - \mathbf{e}_{\text{proj}} \tag{3}$$

$$\mathbf{e}_{\text{emo}} = \alpha \cdot \mathbf{e}_{\text{proj}} + \texttt{detach}(\mathbf{e}_{\text{res}}) \tag{4}$$

where α is learnable (init 2.0). Detaching $\mathbf{e}_{\text{res}}$ in the emotion pathway prevents intensity words ("low," "high") from influencing the base weight direction, enabling disentanglement (Section 3.7).

3.3 Dynamic Graph Attention Network

We model the 8 emotions as nodes in a DGAT (Ding et al., 2025), each node $\mathbf{h}_i \in \mathbb{R}^{512}$ initialized with its CLIP embedding, using multi-head attention with momentum-updated adjacency.

$$e_{ij} = \text{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]) \tag{5}$$

$$w_{ij} \leftarrow \beta \cdot w_{ij} + (1-\beta) \cdot e_{ij} \tag{6}$$

$$\alpha_{ij} = \frac{\exp(w_{ij})}{\sum_k \exp(w_{ik})} \tag{7}$$

$$\mathbf{h}_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_j \alpha_{ij}^k \cdot \mathbf{W}^k \cdot \mathbf{h}_j\right) \tag{8}$$

with $K = 4$ heads, $\beta = 0.5$ for the momentum update (Eq. 6), and a text-conditioned gating mechanism. The momentum adjacency learns inter-emotion relationships (*e.g.*, anger and contempt share high arousal).

3.4 Cross-Attention with Learned Temperature

The feature $\mathbf{e}_{\text{emo}}$ attends to the $N = 8$ DGAT nodes via multi-head cross-attention.

$$\text{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \cdot \tau\right)\mathbf{V} \tag{9}$$

with $\mathbf{Q} = \mathbf{W}_Q \mathbf{e}_{\text{emo}}$, $\mathbf{K} = \mathbf{W}_K \mathbf{H}$, $\mathbf{V} = \mathbf{W}_V \mathbf{H}$ from the node matrix $\mathbf{H} \in \mathbb{R}^{8\times 512}$, learnable temperature τ, $K = 4$ heads, $d_k = 128$, followed by a residual and LayerNorm, $\mathbf{f} =$

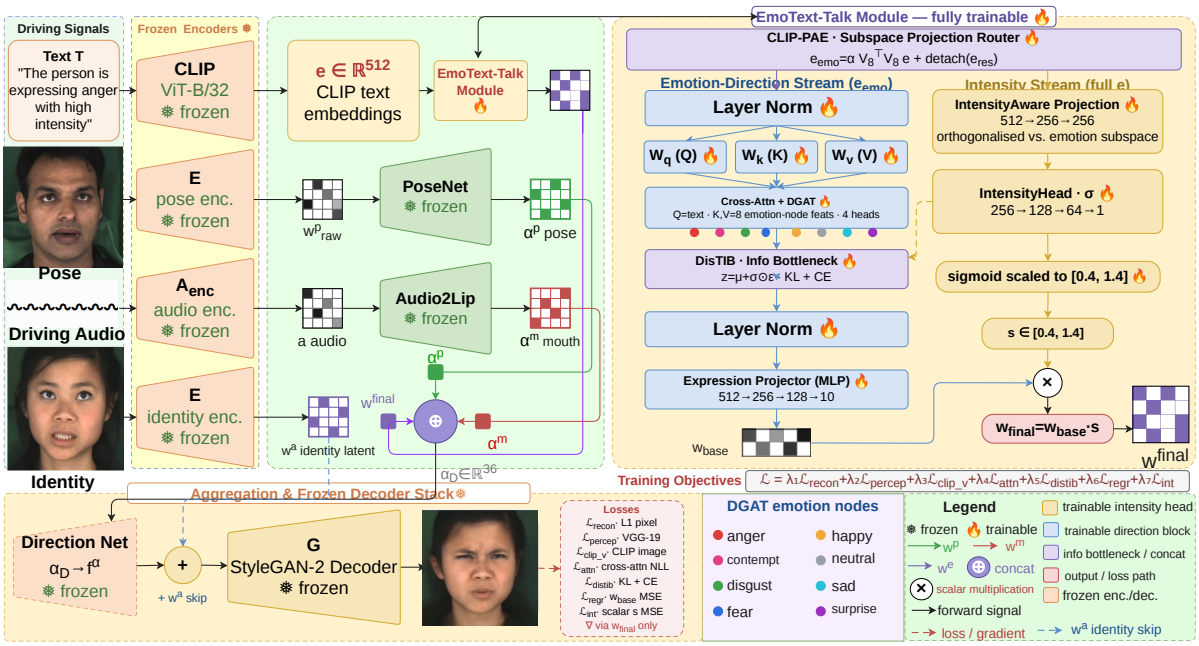

Figure 2: Overview of the proposed EmoText-Talk architecture. We retain EDTalk's frozen encoder **E**, basis banks $\mathbf{B}^m/\mathbf{B}^p$, and StyleGAN-2 decoder **G**, replacing only the expression pathway (ELN) with our text-driven module (blue region). A text prompt T is encoded by CLIP into $\mathbf{e}_T \in \mathbb{R}^{512}$, projected via **CLIP-PAE**, routed through a **DGAT** emotion graph, fused by 4-head **cross-attention**, regularised by **DisTIB**, and decoded to $\mathbf{w}_{\text{base}} \in \mathbb{R}^{10}$. An **IntensityAwareProjection** and **IntensityHead** output scalar s. The final $\mathbf{w}_{\text{final}} = \mathbf{w}_{\text{base}} \cdot s$ is multiplied with $\mathbf{B}^e$ and decoded by **G**. Gradients propagate exclusively through $\mathbf{w}_{\text{final}}$.

$\text{LayerNorm}(\text{CrossAttn}(\cdot)+\mathbf{e}_{\text{emo}})$. Attention supervision (Appendix A) drives τ high, yielding near-one-hot weights $\mathbf{a} \in \mathbb{R}^8$ that route each prompt to the correct node and provide an interpretable visualization.

3.5 Stochastic Encoder (DisTIB)

Following the Transmitted Information Bottleneck (DisTIB) (Dang et al., 2024), we encode $\mathbf{f}$ stochastically ($\boldsymbol{\mu} = \text{MLP}_\mu(\mathbf{f})$, $\log \boldsymbol{\sigma}^2 = \text{MLP}_\sigma(\mathbf{f})$, $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$) with a compression term toward a unit Gaussian and a prediction term retaining emotion information.

$$\mathcal{L}_{\text{distib}} = \underbrace{D_{\text{KL}}(q(\mathbf{z}|\mathbf{f}) \| \mathcal{N}(\mathbf{0}, \mathbf{I}))}_{\text{compression}} - \underbrace{\mathbb{E}[\log p(y|\mathbf{z})]}_{\text{prediction}} \tag{10}$$

where $\beta = 0.01$ balances the terms. Noise injection at training prevents overfitting to the 18 actors. The mean $\boldsymbol{\mu}$ is used at inference for stable outputs on unseen identities.

3.6 Expression Projector

The representation $\mathbf{z} \in \mathbb{R}^{512}$ is projected to EDTalk's 10-D expression weights via a 3-layer MLP, encoding the emotion *direction* independent of intensity.

$$\mathbf{w}_{\text{base}} = \text{MLP}_{512\to256\to128\to10}(\mathbf{z}) \tag{11}$$

3.7 Disentangled Intensity Control

We disentangle *category* from *magnitude* via three mechanisms. **Gradient isolation.** As in Eq. 4, the residual $\mathbf{e}_{\text{res}}$ (intensity words) is detached in the emotion pathway, so the projector learns intensity-invariant weights, and residual values still flow at inference for reconstruction quality. **IntensityAwareProjection.** The full embedding $\mathbf{e}$ is passed through a 2-layer MLP that amplifies the intensity signal.

$$\mathbf{f}_{\text{int}} = \text{ReLU}(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1\mathbf{e}+\mathbf{b}_1)+\mathbf{b}_2) \tag{12}$$

with $\mathbf{f}_{\text{int}} \in \mathbb{R}^{256}$. Raw CLIP embeddings for "low" and "high intensity" have cosine 0.976, and after projection this drops to 0.046 (Table 2), a 95% reduction. **IntensityHead.** A small MLP predicts a continuous scalar

$$s = 0.4+\sigma(\text{MLP}_{256\to128\to64\to1}(\mathbf{f}_{\text{int}})) \in [0.4, 1.4] \tag{13}$$

supervised by MEAD annotations (level_1→0.6, level_2→0.8, level_3→1.0). It correctly orders all

7 non-neutral emotions (mean range 0.183). The final weights are

$$\mathbf{w}_{\text{final}} = \mathbf{w}_{\text{base}} \cdot s \tag{14}$$

3.8 Classifier-Guided Weight Optimization

EDTalk's predefined per-emotion weights were tuned to read as plausible expressions to a human viewer, not to be classified as the intended emotion. Under our held-out protocol, recognition accuracy drops to 0% for contempt and 55% for fear. The cause is structural. EDTalk's expression space is only 10-dimensional and its axes are entangled, so the weight vector a person would call "contempt" and the region a trained recogniser scores as contempt sit at different points in that space, and the hand-set weights land outside the recogniser's high-confidence region for the target emotion. In principle the projector could recover the correct directions on its own from the reconstruction and perceptual losses, but not within a practical training budget. On a talking-head frame, nearly all pixel variation comes from identity, head pose, lighting, and mouth shape, and expression is only a small residual, so the reconstruction gradient is dominated by those factors and its component along the emotion direction is too small to move the weights reliably. With the regression term removed, training collapses to the near-neutral expressions we report. We therefore supply this signal explicitly. For each emotion k we search the 10-D weight space directly for a target that a pretrained classifier reads as k with high confidence, and regress the projector toward it.

$$\mathbf{w}_k^* = \arg\max_{\mathbf{w}} \min_{a \in \mathcal{A}_{\text{train}}} P_\phi(y{=}k \mid G(I_a, \mathbf{w})) \tag{15}$$

where G is the frozen generator, I_a a neutral image of actor a, $\mathcal{A}_{\text{train}}$ the 18 training actors, and P_ϕ a pretrained ResNet-18 classifier. Because the target space is only ten-dimensional, a few hundred classifier-scored random samples (200 iterations per emotion) suffice to locate a strong direction. A gradient-free search also sidesteps the adversarial, classifier-fooling artifacts that gradient ascent on a recogniser tends to produce. The max-min form is what makes the result identity-agnostic which optimizes the worst-case actor rather than the average. This prevents the search from exploiting any single actor's idiosyncratic geometry and forces a weight that transfers to all of them. These targets supervise the projector.

$$\mathcal{L}_{\text{regr}} = \frac{1}{B} \sum_{i=1}^{B} \|\mathbf{w}_{\text{base}}^{(i)} - \mathbf{w}_{k_i}^*\|_2^2 \tag{16}$$

This decouples the two hard problems. The first being *finding* good expression directions, done once and offline with a discriminative classifier, and the second problem of *learning to map text* onto them, which is the projector's task. This is why the approach succeeds where end-to-end reconstruction stalls, and why it is the single largest contributor in our ablation (Section 4.4).

3.9 Integration with EDTalk Generator

The weights $\mathbf{w}_{\text{final}} \in \mathbb{R}^{10}$ are concatenated with lip weights ($\boldsymbol{\alpha}_{\text{lip}} \in \mathbb{R}^{20}$) and pose weights ($\boldsymbol{\alpha}_{\text{pose}} \in \mathbb{R}^{6}$) into a 36-D control vector $\boldsymbol{\alpha}_D = [\boldsymbol{\alpha}_{\text{lip}}; \boldsymbol{\alpha}_{\text{pose}}; \mathbf{w}_{\text{final}}]$, which drives EDTalk's direction network and is decoded by the frozen generator. All EDTalk components stay frozen throughout.

4 Experiments

4.1 Experimental Setup

Dataset. We use MEAD (Wang et al., 2020) (11,905 clips, 18 actors, 8 emotions, 3 intensity levels), holding out 4 actors (W016, W018, M039, M040) for *zero-shot* evaluation. No test actor appears in any training stage, including classifier training and weight optimization.

Implementation. We train for 30 epochs with a batch size of 8, using the AdamW optimizer with a learning rate of 5×10^{-5} under a cosine-annealing schedule. CLIP ViT-B/32 and the full EDTalk generator stay frozen. Weight optimization uses 200 random-search iterations per emotion with a ResNet-18 classifier on the 18 training actors. Training takes ∼6 hours on one RTX 5060 Ti. The module has ∼2.5M trainable parameters (∼2.3M emotion, ∼238K intensity).

Evaluation. We generate 288 test videos (4 actors × 8 emotions × 9 clips), using the standard protocol of EDTalk (Tan et al., 2024) and EmoCAST (Jiang et al., 2025), and report emotion-recognition accuracy Acc_{emo}.

4.2 Comparison with State-of-the-Art

Table 1 compares against existing methods on MEAD. Cross-paper comparison is unreliable due to differing splits, classifiers, and protocols. Sources are marked accordingly.

Method	Control	Acc_{emo}↑
EAMM (Ji et al., 2022)†	video	31.08
PD-FGC (Wang et al., 2023)†	video	44.86
StyleTalk (Ma et al., 2023b)†	video	63.49
EAT (Gan et al., 2023)†	label	64.40
EDTalk-V (Tan et al., 2024)†	video	68.85
EmoHead (Shen et al., 2025)‡	label	75.98
EmoCAST (Jiang et al., 2025)‡	text	83.60
Ours	**text**	**91.67**

Table 1: Comparison with state-of-the-art on MEAD. †From EDTalk (Tan et al., 2024) Table 1. ‡From respective papers. Our result uses the same standard evaluation protocol specified in EDTalk (Tan et al., 2024), EAT (Gan et al., 2023), etc.

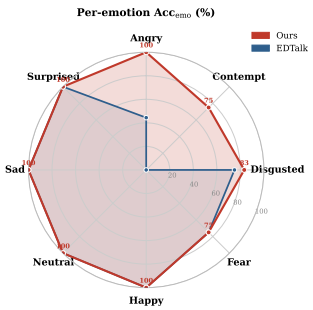

Figure 3: Per-emotion emotion-recognition accuracy (Acc_{emo}, %) on MEAD under the same standard evaluation protocol specified in EDTalk (Tan et al., 2024), EAT (Gan et al., 2023), etc. Our text-driven plug-in dominates the frozen EDTalk baseline on every emotion, with the largest gains on the hardest categories (Contempt, Angry).

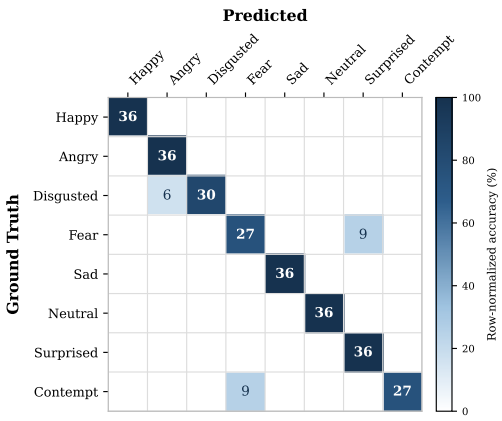

Figure 4: Confusion matrix of EmoText-Talk on the 288 held-out test clips (4 actors × 8 emotions × 9 clips). Rows are ground truth, columns predicted, cells coloured by row-normalised accuracy. Errors are confined to three off-diagonal cells, namely Disgusted→Angry (6), Fear→Surprised (9), and Contempt→Fear (9). The other five emotions are classified perfectly.

EmoText-Talk achieves 91.67% Acc_{emo}, surpassing EDTalk-V (+22.8) and EmoCAST (+8.1) . While EmoCAST trains a full Stable Diffusion pipeline on a 158-hour dataset, we add only ∼2.5M parameters on a frozen generator. Since our module operates exclusively on the expression coefficients while the audio-to-lip pipeline remains entirely frozen, lip synchronization quality is preserved identically to EDTalk's baseline. Reconstruction-fidelity metrics stay close to EDTalk's. Structural similarity (SSIM) is 0.67 versus 0.77 and peak signal-to-noise ratio (PSNR) is 20 dB versus 22 dB. Figure 3 gives the per-emotion breakdown, with perfect accuracy on 5 of 8 emotions (Angry, Happy, Sad, Neutral, Surprised), with the residual errors confined to three off-diagonal cells. As the confusion matrix (Figure 4) shows, Disgusted is occasionally read as Angry (6/36), while Fear and Contempt, which are the two subtlest, lowest-displacement expressions in EDTalk's compact space, are confused with the high-arousal Surprised (9/36) and with Fear (9/36) respectively. The remaining five emotions are never misclassified.

Qualitative comparison. Figure 5 compares methods on a shared source identity and driving audio, so the target emotion is the only variable across rows. Our outputs (bottom row) render each target with clear, well-localised facial cues. The happy targets show raised cheeks and a genuine smile with the lip corners pulled up, while the fearful targets show raised, drawn-together brows, tensed eyes, and a tensed, slightly parted mouth. In both cases the source identity stays intact and the lips remain aligned with the audio. The label- and video-driven baselines ((Gan et al., 2023), (Tan et al., 2024)) tend to produce muted or generic expressions. Since they cannot take free-form text as input, they either express the emotion weakly or fall back to a near-neutral face, and the video-driven variant also leaks appearance cues from its reference clip. (Jiang et al., 2025), a text-driven diffusion baseline that synthesises frames from scratch instead of steering a frozen generator, does reach a visible emotion, but it pays for this with noticeable identity drift and softer detail compared to ours. These visual trends track the quantitative ranking in Table 1 and confirm that the emotion is genuinely rendered on the face, not merely classified correctly.

Identity and lip-sync preservation. Since our method steers a frozen EDTalk (Tan et al., 2024) generator rather than retraining it, the base model's

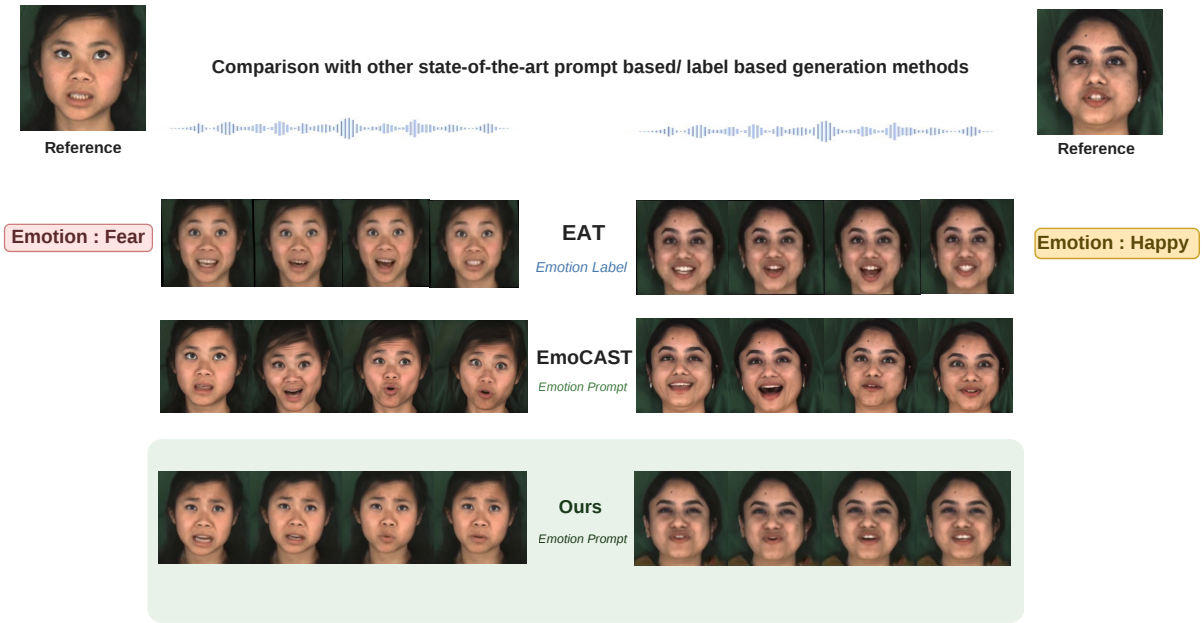

Figure 5: Qualitative comparison with state-of-the-art text-prompt and label-based emotional talking head methods. We compare against representative baselines on the same source identity and driving audio, varying the target emotion across samples. Each row corresponds to a different method, with our method shown in the bottom row.

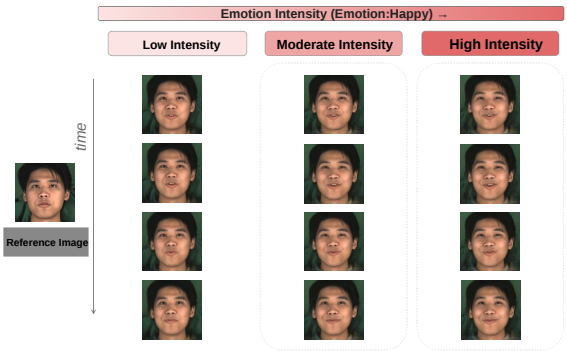

Figure 6: Intensity control results. Same source identity, same emotion prompt ("happiness"), three intensity levels. The learned IntensityHead produces increasingly pronounced expressions from left (low) to right (high).

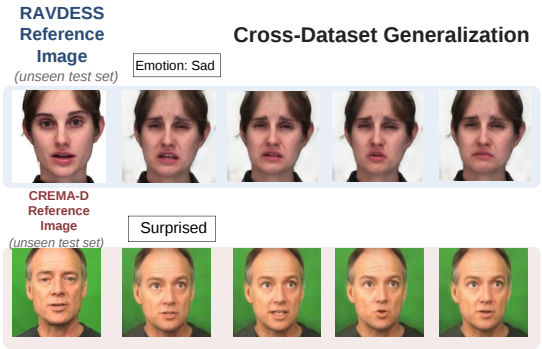

Figure 7: Cross-dataset generalization on RAVDESS and CREMA-D. Without any retraining, our model generates recognizable emotional expressions on unseen identities drawn from both datasets, demonstrating that the learned emotion directions transfer across distinct recording conditions and demographics.

lip-sync and rendering quality largely carry over. The qualitative results confirm this. Our lip articulation tracks the audio as closely as EDTalk's own output, and image quality stays very close to it, with only minor degradation. The added emotion control thus comes at little visible cost to mouth accuracy or sharpness.

4.3 Intensity Control Analysis

Table 2 reports learned intensity values and projection separation. The IntensityHead orders all 7 emotions correctly (mean range 0.183), and the IntensityAwareProjection drops average cosine similarity from 0.976 (raw) to 0.046 (projected), a 95% reduction.

Negative projected cosines for contempt (-0.311) and disgusted (-0.097) push low/high variants to opposite sides of the space, giving maximally discriminable signals. We apply an inference-time linear remapping that is analogous to classifier-free diffusion guidance (Ho and Salimans, 2022) and used to map the learned range to perceptible magnitudes without per-emotion tuning.

Emotion	Low	Mod	High	Range	Raw	Proj
Angry	0.724	0.807	0.907	0.182	0.977	0.075
Happy	0.720	0.813	0.907	0.188	0.976	0.118
Sad	0.725	0.804	0.905	0.180	0.979	0.163
Fear	0.748	0.814	0.903	0.155	0.987	0.256
Contempt	0.724	0.817	0.923	0.199	0.969	−0.311
Disgusted	0.698	0.795	0.912	0.215	0.965	−0.097
Surprised	0.732	0.811	0.897	0.165	0.977	0.317
Average	**0.724**	**0.809**	**0.908**	**0.183**	**0.976**	**0.046**

Table 2: Intensity analysis. "Low/Mod/High" = predicted intensity scalar. "Range" = high−low. "Raw" = CLIP cosine similarity between low/high prompts. "Proj" = cosine after IntensityAwareProjection. All emotions show correct ordinal ordering.

Configuration	$\text{Acc}_{\text{emo}}\uparrow$	Δ
Full model	**91.67**	
− DisTIB (λ_3=0)	90.63	−1.04
− IntensityHead	89.24	−2.43
− Attn. supervision (λ_5=0)	77.08	−14.59
− CLIP-PAE (α=0)	29.17	−62.50
− Regression (λ_6=0)	29.17	−62.50

Table 3: Ablation study. Each row removes one component from the full model. Acc_{emo} uses the same standard evaluation protocol specified in EDTalk (Tan et al., 2024), EAT (Gan et al., 2023), etc. Δ = drop from full model.

4.4 Ablation Study

Table 3 ablates one component at a time. Each variant is retrained with identical hyperparameters and evaluated on 288 clips.

A three-tier hierarchy emerges. **Tier 1 (foundations).** Removing either CLIP-PAE or regression collapses accuracy to 29.17% (−62.50), and the failures are complementary, that is, without CLIP-PAE prompts collapse into the Fear and Contempt basins, while without regression, outputs flatten toward neutral. **Tier 2 (routing).** Attention supervision contributes −14.59, as the model loses the ability to route fine-grained emotion distinctions through the graph attention. **Tier 3 (calibration).** The IntensityHead adds +2.43, with the largest effect on the subtle low-displacement emotions (Contempt, Fear), and also enables text-driven intensity control. DisTIB adds +1.04 via stochastic regularization.

4.5 Cross-Dataset Generalization

We apply our model to the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo, 2018) (24 studio actors, 8 emotions) and the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) (Cao et al., 2014) (91 actors, diverse ethnicities, 6 emotions, naturalistic conditions) *without retraining*. Both differ from MEAD in lighting, viewpoint, and demographics. After alignment to 256×256, we use these faces as identity sources, drive them with MEAD audio, and control expression via text. As shown in Figure 7, our model produces recognizable expressions on both despite never seeing these faces, confirming it operates in abstract expression space rather than memorizing actor-specific configurations.

5 Conclusion

We presented EmoText-Talk, a lightweight plug-in (∼2.5M parameters) for text-driven emotion and intensity control. Combining CLIP-PAE projection, dynamic graph attention routing, classifier-guided regression, information-bottleneck encoding, and disentangled intensity prediction, it achieves 91.67% emotion accuracy on MEAD, surpassing EDTalk-V (+22.8) and EmoCAST (83.60%) with no inference overhead. The IntensityAwareProjection cuts intensity-level CLIP cosine similarity from 0.976 to 0.046, giving ordinal control for all 7 emotions, and ablations confirm CLIP-PAE and classifier-guided regression are each worth over 60 points. Natural language thus offers a flexible interface for facial expression control.

Limitations

Fixed vocabulary. The 8-node DGAT routes novel emotions ("confused," "sarcastic") to the nearest

node. Softening temperature enables blending but expressions stay bounded by 8 nodes. Open-vocabulary control via valence-arousal-dominance (VAD) spaces (Cho et al., 2024) or text-encoder fine-tuning (Zhong et al., 2024) is future work. **Intensity range.** The visual gap between low and high is subtle (range 0.183) due to EDTalk's compact space. We use inference-time remapping, but ordinal contrastive losses (Zha et al., 2023) during training could widen it. **Generator specificity.** Targets and scaling are tuned for EDTalk's 10-D space. Adapting to diffusion generators requires re-optimizing them, though the framework is architecture-agnostic. **CLIP limitations.** The encoder inherits bag-of-words behavior (Koishigarina et al., 2026) and weak negation handling. Our projections mitigate this but nuanced descriptions would benefit from fine-tuned encoders. **Prompt understanding.** Trained on structured prompts, the model may not map cleanly from free-form input. A lightweight LLM front-end could rewrite arbitrary input into the canonical format. **Head pose and gaze.** We modify only expression weights, leaving pose, blinking, and gaze to EDTalk's frozen pathway. Extending text conditioning to the pose network would yield more holistically expressive results. **Inherited visual quality.** Because we tweak only the expression weights of a frozen EDTalk backbone, our output fidelity is inherited from and bounded by EDTalk. Visual quality (SSIM, PSNR) stays close to EDTalk's reported values, and any artifacts of the base generator carry through unchanged. Our method cannot improve resolution, texture detail, or temporal stability beyond what EDTalk produces. Gains there would require a stronger backbone, to which our module could be re-targeted.

Ethics Statement

EmoText-Talk synthesises emotionally expressive talking-head videos from a single reference identity, and like all face-generation technology it carries a risk of misuse, including non-consensual impersonation, misinformation, and the fabrication of misleading affective content. We release this work for research purposes and advocate its responsible use. Synthesised videos should be clearly disclosed as artificial, produced only with the consent of the depicted individual, and accompanied by provenance or watermarking safeguards where appropriate. Our experiments use only publicly available, consent-based research datasets (MEAD (Wang et al., 2020), RAVDESS (Livingstone and Russo, 2018), and CREMA-D (Cao et al., 2014)) recorded from actors who agreed to the distribution of their performances for academic study, and we use them solely under their respective licences. Because our module steers a frozen, pretrained generator rather than training a new one, it introduces no additional identity data beyond these public corpora. We encourage practitioners deploying such systems to pair them with reliable synthetic-media detection and explicit labelling in order to mitigate potential downstream harms.

Use of AI Assistants

AI-based writing assistants were used to improve the language of this manuscript, including its grammar, clarity, and phrasing. All technical content, including the method, experiments, analyses, and claims, is the authors' own, and the authors reviewed and take full responsibility for the final text.

Code Availability

Our code is available at `https://github.com/noidentitysubmit45-afk/EmoText-Talk`.

References

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. 2024. EmoSphere-TTS: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. In *Proceedings of Interspeech*.

Zhuohang Dang, Minnan Luo, Chengyou Jia, Guang Dai, Jihong Wang, Xiaojun Chang, and Jingdong Wang. 2024. Disentangled representation learning with transmitted information bottleneck. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*.

Shihang Ding, Kaixuan Wang, Wenhao Jiang, Cong Xu, Hongjian Bo, Lin Ma, and Haifeng Li. 2025. DGAT: A dynamic graph attention neural network framework for EEG emotion recognition. *Frontiers in Psychiatry*, 16.

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. 2023. Efficient emotional adaptation

for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. 2022. EAMM: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*.

Yiguo Jiang, Xiaodong Cun, Yong Zhang, Yudian Zheng, Fan Tang, and Chi-Man Pun. 2025. EmoCAST: Emotional talking portrait via emotive text description. *arXiv preprint arXiv:2508.20615*.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. 2026. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson audio-visual database of emotional speech and song (RAVDESS). *PLoS ONE*, 13(5).

Yifeng Ma, Suzhen Wang, Yu Ding, Lincheng Li, Bowen Ma, Tangjie Lv, Changjie Fan, Zhipeng Hu, Zhidong Deng, and Xin Yu. 2023a. TalkCLIP: Talking head generation with text-guided expressive speaking styles. *arXiv preprint arXiv:2304.00334*.

Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. 2023b. StyleTalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Xuli Shen, Hua Cai, Dingding Yu, Weilin Shen, Qing Xu, and Xiangyang Xue. 2025. EmoHead: Emotional talking head via manipulating semantic expression parameters. *arXiv preprint arXiv:2503.19416*.

Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. 2024. EDTalk: Efficient disentanglement for emotional talking head synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. 2024. EMO: Emote portrait alive — generating expressive portrait videos with audio2video diffusion model under weak conditions. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. 2023. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yibo Xia, Lizhen Wang, Xiang Deng, Xiaoyan Luo, and Yebin Liu. 2024. GMTalker: Gaussian mixture-based audio-driven emotional talking video portraits. *arXiv preprint arXiv:2312.07669*.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. 2024. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*.

Chaolong Yang, Kai Yao, Yuyao Yan, Chenru Jiang, Weiguang Zhao, Jie Sun, Guangliang Cheng, Yifei Zhang, Bin Dong, and Kaizhu Huang. 2025. Unlock pose diversity: Accurate and efficient implicit keypoint-based spatiotemporal diffusion for audio-driven talking portrait. *arXiv preprint arXiv:2503.12963*.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. 2023. Rank-N-Contrast: Learning continuous representations for regression. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jian Zhang, Weijian Mai, and Zhijun Zhang. 2024. EMOdiffhead: Continuously emotional control in talking head generation via diffusion. *arXiv preprint arXiv:2409.07255*.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. 2024. MEMO: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*.

Yicheng Zhong, Huawei Wei, Peiji Yang, and Zhisheng Wang. 2024. ExpCLIP: Bridging text and facial expressions via semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Chenliang Zhou, Fangcheng Zhong, and Cengiz Öztireli. 2023. CLIP-PAE: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable, and controllable text-guided face manipulation. In *ACM SIGGRAPH 2023 Conference Proceedings*.

A Training Objectives

We train end-to-end with 7 losses, $\mathcal{L} = \sum_{i=1}^{7} \lambda_i \mathcal{L}_i$. **Reconstruction** ($\lambda_1$=1.0), $\|I_{\text{gen}} - I_{\text{gt}}\|_1$. **Perceptual** ($\lambda_2$=0.5), VGG-19 matching $\sum_l \|\phi_l(I_{\text{gen}}) - \phi_l(I_{\text{gt}})\|_2^2$. **DisTIB** ($\lambda_3$=0.1), Eq. 10. **CLIP verification** (λ_4=0.1), $1-\cos(\text{CLIP}_{\text{img}}(I_{\text{gen}}), \mathbf{e})$. **Attention supervision** (λ_5=1.0), $-\log \mathbf{a}_y$, routing emotions to the correct node. **Regression** (λ_6=0.5), Eq. 16. **Intensity** (λ_7=0.5), $\|s - s_{\text{gt}}\|_2^2$.