SCHRONO: CONDITIONAL DIFFUSION MODELING VIA SENSORY-CHRONO HIERARCHY FOR MULTI-MODAL COUNSELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal counseling integrates visual, acoustic, and textual cues to understand a client's emotions and provide empathetic support. However, two architectural weaknesses remain. First, the two objectives are coupled: perception updates perturb the distribution of generated text. Second, causal attention restricts each video frame to earlier frames, so a late disclosure cannot revise an early reading. We introduce SCHRONO, a framework for multimodal emotion understanding and counseling response generation. It uses a frozen masked-diffusion language model with two components: the Sensory layer trains encoders, projectors, attention pooling, and light heads for emotion and strategy prediction; the Chrono layer takes video, audio, and dialogue history to generate counseling responses through masked diffusion. Separate trainable pathways decouple perception from generation. Bidirectional attention lets each frame access the full video and dialogue context. On the MESC dataset, SCHRONO outperforms the compared methods in emotion recognition and counseling strategy prediction. Its responses also receive the highest average trustworthiness ratings from LLM judges. Our method also recognizes emotions across datasets and generates coherent supportive responses on ESConv without training on that corpus. Further fine-tuning improves 13.1 points for emotion recognition on IEMOCAP. Replacing bidirectional attention with causal attention in the same model lowers response quality, supporting the value of full-context access.

1 INTRODUCTION

Mental health disorders affect one in eight people worldwide, yet access to professional counseling is limited by cost, availability, and stigma, making AI-assisted counseling a valuable target. Counseling dialogues involve not only words but also facial expressions and tone of voice. Clients rarely disclose distress directly, since surface symptoms such as insomnia or low mood mask deeper concerns that emerge only later (Liu et al., 2021; Chu et al., 2025; Cheng et al., 2026), and the therapist selects among techniques such as restatement and interpretation while reading facial expression, prosody, and wording at once. A system for this setting must perceive all modalities without committing to an interpretation early: a premature commitment makes clients feel unheard, and ignoring video or audio misses the emotional state carried outside the transcript.

The dominant alternative adapts a causal autoregressive language model by supervised fine-tuning (SFT), projecting video and audio features into the token embedding space and training perception and generation jointly (Cheng et al., 2024b; Le et al., 2026). It has two architectural weaknesses: it couples the two objectives, so updating the backbone for perception changes the distribution of its generated text, and its causal attention restricts each frame to earlier frames, so a late disclosure cannot revise an early reading.

We therefore build SCHRONO on a masked-diffusion language model (Austin et al., 2021; Li et al., 2022; Sahoo et al., 2024; Nie et al., 2025) whose backbone is kept frozen, training only the perception machinery around it. Bidirectional attention gives every token, including each injected video frame, global visibility over the whole video and dialogue, matching progressive disclosure, while freezing the backbone ensures that perception training leaves the pretrained generator weights

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.

untouched. We realize this as a Sensory-Chrono hierarchy that shares one backbone between a discriminative and a generative layer (Figure 1); because the two layers occupy disjoint parameters, the same model is at once an accurate recognizer and an intact generator.

To separate the effect of information-flow topology from differences between pretrained backbones, we evaluate the bidirectional design at two levels (Section 4.5): a comparison of LLaDA-8B and Llama-3.1-8B under matched conditioning recipes, and a same-checkpoint intervention that changes only the attention mask.

Our contributions are as follows:

- We introduce SCHRONO, a Sensory-Chrono framework that decouples perception and generation through separate trainable pathways around a frozen diffusion backbone.
- We analyze bidirectional conditioning through cross-backbone comparisons and a same-checkpoint mask intervention, which support the role of bidirectional attention in improving response quality.
- Our method improves emotion recognition, counseling strategy prediction, and supportive response quality, achieving leading results on MESC with evidence of transfer across datasets.

2 Related Work

Counseling systems and multimodal perception. Counseling systems for mental health build on psychotherapy frameworks rather than generic chit-chat, with client-centered and cognitive-behavioral principles underlying most AI counselors (Rogers, 1957; Beck et al., 1979); PsyQA (Sun et al., 2021), CPsyCoun (Zhang et al., 2024), and PsyDT (Xie et al., 2025) are text-only systems that fine-tune a language model. Emotional-support dialogue is where the evaluation practice and multimodal data come from: text-based support (Liu et al., 2021) established dialogue-level evaluation of empathetic responses, its multimodal extension (Chu et al., 2025) added video and audio, MultiMood (Le et al., 2026) built a large multimodal counseling model with a seven-dimension trustworthiness rubric, and graph-enhanced multimodal LLMs (Xiong et al., 2026) and BiGraph-Diffuse (Cheng et al., 2026) extend this to hierarchical emotion understanding and masked-diffusion counseling respectively. All of these fine-tune the underlying language model, coupling perception and generation on shared backbone gradients. On the recognition side, emotion recognition in conversation has progressed from recurrent and graph architectures (Majumder et al., 2019; Ghosal et al., 2019) to pretrained face (Zhai et al., 2023; Mollahosseini et al., 2019) and speech encoders (Chen et al., 2022; Poria et al., 2019); these excel at classification but do not generate supportive responses (Appendix A). Our setting differs from both lines: we target multimodal therapeutic counseling for mental-health-related distress, coded with therapy techniques rather than generic empathy and progressing through staged disclosure. Rather than fine-tuning, SCHRONO keeps a bidirectional masked-diffusion backbone frozen and trains only the perception machinery around it, so recognition and generation occupy disjoint parameters and neither can damage the other.

Diffusion language models and efficient adaptation. Discrete diffusion (Austin et al., 2021; Li et al., 2022) models text with a corruption-and-reconstruction objective, and masked-diffusion variants (Sahoo et al., 2024; Nie et al., 2025) scale this to billions of parameters; LLaDA (Nie et al., 2025) shows that an 8B masked-diffusion transformer matches autoregressive baselines on in-context learning while retaining the bidirectional topology, and classifier-free guidance (Ho & Salimans, 2022) applies naturally to such models. Multimodal adaptation of these backbones is early, and Appendix A details the closest work (LLaDA-V, DIFFA, LLaDA-o, LLaDA2.0-Uni) and why each occupies a different point in the design space. Adapters that keep a backbone frozen, such as LoRA (Hu et al., 2022), prefix tuning (Li & Liang, 2021), and prompt tuning (Lester et al., 2021), were developed for causal LLMs, and multimodal models such as LLaVA (Liu et al., 2023) followed the same recipe, with prefix-level learning signals also studied in reward-directed adaptation (Wang et al., 2026b). Such variants typically still train the projector jointly with the backbone or adapter and inherit the causal information flow. SCHRONO freezes the backbone completely and trains only encoders, projectors, pooling, prefixes, and heads, making the adaptation strictly additive; attention pooling over frame features (Lee et al., 2019) replaces mean pooling so that short emotional events are not averaged away.

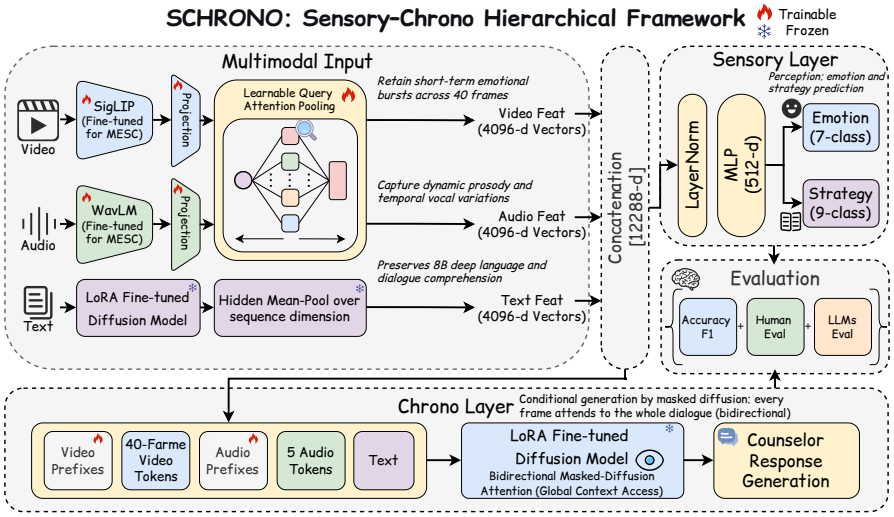

Figure 1: The SCHRONO architecture. SigLIP, WavLM, and the BiGraph-Diffuse base encode video, audio, and text. The discriminative Sensory layer pools and concatenates the three 4096-dimensional representations and predicts the client's emotion (7 classes) and the counselor's strategy (9 classes); the generative Chrono layer feeds the learned prefixes, 40 frame tokens, 5 audio tokens, and dialogue text to the frozen bidirectional masked-diffusion backbone.

3 METHODOLOGY: SCHRONO

3.1 PRELIMINARIES

Masked-diffusion language models. A masked-diffusion language model (Nie et al., 2025; Sahoo et al., 2024) is trained to reconstruct tokens masked by a random binary mask $\boldsymbol{m} \sim \mathrm{Bern}(t)$ with masking ratio $t \sim U(\epsilon, 1)$, predicting them from the observed tokens with a bidirectional transformer (Vaswani et al., 2017), the same all-pairs visibility that makes masked token prediction effective in pretrained language models (Devlin et al., 2019). Following LLaDA's answer-length normalization, the supervised loss for a dialogue with answer length N_b^{ans} is $\mathcal{L}_{\mathrm{gen}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_b^{\mathrm{ans}}}\sum_{i\in\mathcal{M}_b}\frac{-\log p_\theta(x_{b,i}|\boldsymbol{x}_b^{\mathrm{mask}})}{\hat{p}_i}$, over the masked positions $\mathcal{M}_b$ of sample b and their masking probabilities $\hat{p}_i$. This objective minimizes an upper bound on the negative log-likelihood of the reverse model's generative distribution (statement and derivation in Appendix B). Generation proceeds by iterative unmasking: at each step the model scores all masked positions and reveals the most confident first, with Gumbel noise for sampling (Nie et al., 2025).

Task setup. MESC (Chu et al., 2025; Le et al., 2026) is a multimodal emotional support conversation benchmark. Each training example is a dialogue context C of the last K turns with the corresponding client video and audio clips, and a target therapist turn. We address user emotion recognition (7 classes), therapist strategy prediction (9 classes), and generation of the therapist response. All three receive only the dialogue context as input: the gold labels are used exclusively as supervision and never appear in the input. Video clips are sampled to 8 frames each, so with $K = 5$ context turns the video pathway contributes $5 \times 8 = 40$ per-frame tokens and the audio pathway 5 tokens; Appendix C compares this tokenization with the MultiMood pipeline.

3.2 SENSORY LAYER

Video pathway. Each frame is encoded by SigLIP-So400M (Zhai et al., 2023), whose class token is mapped to the LLaDA embedding space by a two-layer projection pretrained on AffectNet (Mollahosseini et al., 2019) and then adapted to MESC in Stage 1. Because a single pooled vector would average out a 1 to 2 second emotional burst, per-frame features $\boldsymbol{F} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_J] \in \mathbb{R}^{J\times D}$ are kept as separate tokens for generation and aggregated by learnable attention pooling (Lee et al., 2019) for classification, $\boldsymbol{z} = \sum_j w_j \boldsymbol{f}_j$ with $w = \mathrm{softmax}(\boldsymbol{F}\boldsymbol{q})$ and $\boldsymbol{q}$ initialized near zero so that pooling warms up from mean pooling.

Audio and text pathways. WavLM (Chen et al., 2022) encodes each audio clip into frame-level features, mapped by a projector pretrained on MELD (Poria et al., 2019) and averaged over time into a single vector per clip; the classification head additionally pools the five clip vectors with learnable attention. The dialogue history is formatted as a structured context with Problem/Situation headers followed by alternating patient and counselor turns; the frozen LLaDA-8B backbone (Nie et al., 2025), with the BiGraph-Diffuse adapter merged (Cheng et al., 2026), encodes it and the last hidden states are mean-pooled. The context stops before the target therapist response, so the text representation never encodes the label or the target response. Each modality is thus summarized into one representation by encoder, projector, and pooling, $\boldsymbol{f}^{\mathrm{m}} = \mathrm{Pool}(\mathrm{Proj}_m(\mathrm{Enc}_m(\mathcal{X}_m)))$ for $m \in \{\mathrm{v}, \mathrm{a}\}$ and $\boldsymbol{f}^{\mathrm{t}} = \mathrm{MeanPool}(\mathrm{LLaDA}(\mathcal{T}))$; feature widths and preprocessing are in Appendix C.

Classification. The three pooled representations $\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{t}$ are concatenated to a $3D = 12288$-dimensional vector, normalized by LayerNorm, and mapped by an MLP with one hidden layer of 512 units and dropout,

$$\boldsymbol{z} = \mathrm{MLP}\big(\mathrm{LN}\big([\boldsymbol{f}^{\mathrm{v}}; \boldsymbol{f}^{\mathrm{a}}; \boldsymbol{f}^{\mathrm{t}}]\big)\big), \tag{1}$$

whose two heads predict the 7-class emotion and the 9-class strategy. The emotion head uses cross-entropy with inverse-frequency class weights $\boldsymbol{w}_e$ and label smoothing ϵ. The strategy head uses focal loss (Lin et al., 2017) with $\gamma = 2$ and the same class weights,

$$\mathcal{L}_{\mathrm{emo}} = \mathrm{CE}(\boldsymbol{z}_e, y_e; \boldsymbol{w}_e, \epsilon), \qquad \mathcal{L}_{\mathrm{strat}} = \sum_c w_{s,c}\,(1 - p_c)^{\gamma}\,\mathrm{CE}_c, \tag{2}$$

where p_c is the predicted probability of the true class c. The sensory loss combines the two terms,

$$\mathcal{L}_{\mathrm{cls}} = \mathcal{L}_{\mathrm{emo}} + \lambda_s\,\mathcal{L}_{\mathrm{strat}}, \tag{3}$$

and for prediction we average the softmax of the multimodal head with a text-only head trained on $\boldsymbol{t}$ alone (late fusion with weight 0.5),

$$p(y) = \tfrac{1}{2}\,\mathrm{softmax}(\boldsymbol{z}_{\mathrm{mm}}) + \tfrac{1}{2}\,\mathrm{softmax}(\boldsymbol{z}_{\mathrm{txt}}). \tag{4}$$

3.3 Chrono layer

The Chrono layer attaches multimodal input to the same frozen backbone used by the Sensory layer. Two learnable prefix tokens are prepended to the video stream and two to the audio stream. They are trained in Stage 2 so that the multimodal features align with LLaDA's text distribution. The full conditioning sequence is

$$[\ \boldsymbol{q}_1^{\mathrm{v}}, \boldsymbol{q}_2^{\mathrm{v}},\ \boldsymbol{f}_1^{\mathrm{v}}, \ldots, \boldsymbol{f}_{40}^{\mathrm{v}},\ \boldsymbol{q}_1^{\mathrm{a}}, \boldsymbol{q}_2^{\mathrm{a}},\ \boldsymbol{f}_1^{\mathrm{a}}, \ldots, \boldsymbol{f}_5^{\mathrm{a}},\ \boldsymbol{x}^{\mathrm{text}}\ ], \tag{5}$$

where each video frame is an independent token in equation 5. The whole sequence passes through the frozen backbone under bidirectional attention, so every frame can attend to all other frames and to the dialogue. Conditioning on the concatenated embeddings of equation 5, the diffusion sampler iteratively predicts and reveals tokens, unmasking the highest-confidence positions first under LLaDA's remasking schedule (Nie et al., 2025), with classifier-free guidance (Ho & Salimans, 2022) interpolating the conditional and zeroed-prefix passes; sampler equations and generation settings are in Appendix C. Because the backbone is frozen, neither classification objective can interfere with its pretrained generation ability.

Why this design suits multimodal counseling. Freezing the backbone leaves the pretrained generator weights untouched, so the perceptual objectives cannot erode the model's ability to produce fluent, empathetic text. Bidirectional attention matches the non-linear nature of counseling: every frame token attends to all other frames and to the dialogue, so a late disclosure can revise the reading of an early frame, and the attention weights rather than a positional constraint decide how much each modality contributes. Section 4.5 quantifies these differences against a causal counterpart with identical inputs, and Appendix B gives the supporting argument.

3.4 Training curriculum

Stage 0 (projector pretraining). The video projector is trained on AffectNet and the audio projector on MELD with classification objectives, initializing emotion-aware sensory projectors.

Stage 1 (sensory adaptation). The SigLIP and WavLM encoders are unfrozen and trained with the projectors, attention pooling, and classifier heads on the MESC classification objective, with mixup and light data augmentation; the projectors are re-initialized from the pretrained weights. The final epoch's weights are the deliverable, and the test set never participates in training or checkpoint selection.

Stage 2 (joint perception-generation). The encoders are frozen. The projectors, learnable prefixes, and classifier heads are trained jointly by minimizing the weighted combination of the generation and classification losses,

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_e \, \mathcal{L}_{\text{emo}} + \lambda_s \, \mathcal{L}_{\text{strat}}, \tag{6}$$

with $\lambda_s = 2.0$ to balance the numerically smaller focal strategy loss against the emotion loss. To keep the two objectives from perturbing each other, the classification pathway uses frozen copies of the Stage 1 projectors. This decoupling is exact: with the backbone frozen and the classification features produced by frozen projector clones, the generation and classification gradients vanish on each other's parameter blocks (Proposition B.1, proof in Appendix B). Cached backbone features make each epoch a lightweight pass over the 8B model, and all reported classification numbers are read out from this Stage 2 model.

Stage 3 (inference). For classification, the heads are applied directly; for generation, the learned prefixes and per-frame tokens are recomputed from the client's video and audio and passed to the masked-diffusion sampler. For controlled analysis, we also define a topology intervention that replaces the bidirectional attention mask of the frozen backbone with a causal mask without modifying any backbone parameter or learned conditioning input. It is used only for the diagnostics in Section 4.5.

4 Experimention

4.1 Experimental Setting

Dataset. We use the official MESC splits (Chu et al., 2025; Le et al., 2026): 1,019 dialogues and 28,762 utterances in total, with a video and an audio clip for every utterance (Chu et al., 2025). The split statistics and the distribution of presenting problems are in Appendix D.

Metrics. User emotion recognition and strategy prediction are measured with accuracy and F1. Generated responses are scored by LLM judges on the seven MultiMood trustworthiness dimensions, each rated from 1 to 10 (Le et al., 2026) (rubric in Appendix E). **Baselines.** We compare against the models reported in the MultiMood benchmark (Le et al., 2026) under the same protocol: multimodal SFT models, video LLMs (Cheng et al., 2024b; Lin et al., 2024; Wang et al., 2024; Cheng et al., 2024a), general LLMs (Yang et al., 2024; Meta AI, 2025; Guo et al., 2025; OpenAI, 2024), and MultiMood's SFT, SFT+GRPO, and compressor variants; Table 2 lists all 23 rows (text-only baselines in Appendix H). **Implementation.** We use PyTorch with bfloat16, initialize the encoders from SigLIP-So400M and WavLM, precompute text features with the frozen LLaDA, and use LLaDA-8B with the merged BiGraph-Diffuse counseling adapter (Cheng et al., 2026) as the backbone. Training runs on a single NVIDIA RTX 6000D GPU in about 24 GPU-hours; all hyperparameters are in Appendix C. For strategy prediction, we predict among the 9 annotated strategies, excluding an undefined catch-all label.

Provenance of the frozen adapter. The adapter is the one component of the backbone we do not train, so its training corpus bounds how much of the generation result can be attributed to the perception stack. BiGraph-Diffuse fine-tunes LoRA adapters on the English OpenR1-Psy (18,852 dialogues) and the Chinese CPsyCounD (3,134 dialogues); neither is MESC, MELD, or ESConv, and our MESC test split is disjoint from its training data. Both corpora are counseling dialogues, so a strong adapter should be expected to help generation on any counseling benchmark, and the informative comparison for our generation claim is therefore SCHRONO against its own variants (Table 22) rather than against MultiMood, which fine-tunes on MESC directly.

4.2 Main results on generation

This is the primary claim of the paper: because the backbone is never updated, adding a perception stack should condition generation without eroding it. We evaluate the generated counselor responses

Table 1: LLM-judge trustworthiness scores (1 to 10) for generated counselor responses, averaged over the seven dimensions. Full per-dimension tables are in Appendix F.

System	GPT-4o	Claude-4.0-Sonnet	Grok-3	Mean
Qwen2-7B	5.01	6.16	6.20	5.79
Llama-3.1-8B (w/o prefix)	5.47	6.21	6.31	6.00
Llama-3.1-8B (w/ prefix)	5.13	6.43	6.52	6.03
MultiMood (SFT+RL)	6.63	7.01	7.81	7.15
SCHRONO (w/o prefix)	8.23	7.36	8.20	7.93
SCHRONO	**8.30**	**7.50**	**8.29**	**8.03**

with three LLM judges (GPT-4o, Claude-4.0-Sonnet, Grok-3) on the complete MESC test responses (454 per system) under the seven-dimension MultiMood
ing baselines, were scored under the same protocol.

We avoid reference-based metrics such as BLEU or ROUGE, which undervalue open-ended empathetic phrasings. SCHRONO receives the highest average score from all three judges (8.30, 7.50, 8.29) and the highest overall mean (8.03), ahead of MultiMood SFT+RL (7.15); the learned prefix contributes to this result (7.93 without it). The advantage is on the average rather than on every dimension: MultiMood SFT+RL is ahead on four of the twenty-one judge-dimension cells, namely empathy (7.57 vs. 7.56) and reliability (7.09 vs. 6.87) under Claude, and fairness (8.61 vs. 7.29) and privacy (8.97 vs. 8.85) under Grok. The full per-dimension tables are in Appendix F, and qualitative examples in Appendix G.

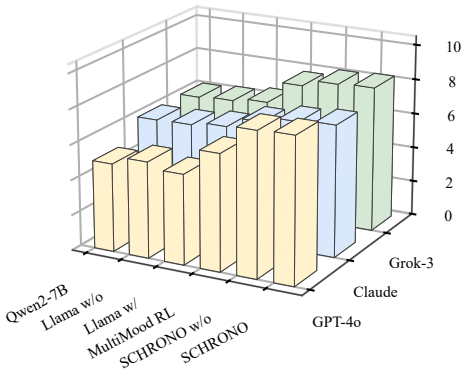

Figure 2: LLM-judge trustworthiness scores (1 to 10), by judge. SCHRONO receives the highest mean score from every judge; exact values in Table 1.

4.3 Main results on emotion and strategy classification

Classification verifies the other half of the decoupling claim: the same frozen generator is conditioned by a perception stack strong enough to be a competitive recognizer, so separation does not come at the cost of perceptual accuracy. Table 2 reports user emotion and strategy prediction as means over three runs (final-epoch weights) on the full test set, under the MultiMood benchmark protocol; the text-only general LLMs from the same benchmark are in Appendix H, since they see only the transcript. SCHRONO reaches 68.70 ± 0.9 accuracy and 66.63 ± 1.0 F1 on user emotion and 50.52 ± 1.1 accuracy and 49.64 ± 1.2 F1 on strategy prediction, the best reported results on all four metrics: against MultiMood SFT+GRPO it improves by 10.10 (emotion accuracy), 8.85 (emotion F1), 7.71 (strategy accuracy), and 9.99 (strategy F1) points, and against SMES, the strongest prior strategy model, by 1.52 points on strategy accuracy.

4.4 Which component of the conditioning stack drives generation?

Freezing the backbone guarantees that perception training cannot degrade the generator; the question this section answers is the complementary one, whether a frozen backbone can still be conditioned effectively, and which part of the stack is responsible. On the MESC test responses, adding the learned multimodal prefix improves the LLM-judge mean from 7.93 to 8.03, so perception-conditioned generation is not worse than base generation. The same frozen base also generates coherent supporter responses on ESConv (Liu et al., 2021), an unseen text-only corpus, scored by the same three judges over 500 supporter turns against Llama-3.1-8B (Grattafiori et al., 2024) and GPT-4o. SCHRONO receives the highest mean from all three judges (8.10, 7.30, 8.09) and the prefix improves the no-prefix variant on every judge; the GPT-4o generator is ahead on a single cell, ethics under the GPT-4o judge (8.51 vs. 7.81), and the GPT-4o row should be read with care throughout, since that judge tends to favor GPT-4o generations. Per-dimension tables are in Appendix I. Table 22 answers the attribution question directly by decomposing the generation result into the frozen adapter, the learned prefix, and the sensory conditioning, all under the same weights.

Table 2: MESC benchmark results for user emotion (7 classes) and strategy prediction (9 classes). A: audio, V: video, T: text. Comp. denotes MultiMood's conversation compressor. Baselines from the MultiMood benchmark (Le et al., 2026); SCHRONO's row: mean ± std over three runs (final-epoch weights) on the full test set. Bold: best in column.

Model	Training	Mod.	User emotion		Strategy prediction	
			Acc	F1	Acc	F1
MMGCN	SFT	A,V,T	55.80	57.58	–	–
MMDFN	SFT	A,V,T	58.13	55.86	–	–
Blenderbot SFT	SFT	A,V,T	–	–	48.00	46.10
SMES	SFT	A,V,T	54.60	46.80	49.00	20.20
VideoLLaMA2-72B	–	A,V,T	55.06	55.68	25.77	26.09
VideoLLaVA	SFT	V,T	46.60	47.08	27.31	22.28
InternVideo2.5-8B	SFT	V,T	37.22	34.69	13.44	4.82
VideoLLaMA3-7B	SFT	A,V,T	45.28	46.23	33.96	24.50
EmotionLLaMA	SFT	A,V,T	46.12	41.95	37.44	25.41
LLaMA4-Maverick	–	T	23.34	21.16	14.53	8.11
Claude-3.7-Sonnet	–	T	32.59	33.33	27.97	27.55
DeepSeek-R1	–	T	20.48	20.27	17.84	15.95
GPT-4o	–	T	38.98	43.56	24.88	26.26
Grok-2	–	T	22.46	25.08	20.44	18.19
Qwen2-7B	SFT	A,V,T	41.83	37.16	37.43	33.52
Qwen2-0.5B	SFT+PPO	A,V,T	44.27	44.99	36.34	33.01
Qwen2-7B	SFT+Comp.	A,V,T	44.27	44.03	39.42	35.69
Qwen2.5-7B	SFT	A,V,T	53.00	51.13	35.14	34.46
MultiMood	SFT+Comp.	A,V,T	53.75	51.75	39.29	36.25
MultiMood	SFT	A,V,T	56.38	55.81	36.78	34.32
MultiMood	SFT+GRPO	A,V,T	58.60	57.78	42.81	39.65
MultiMood	SFT+Comp+GRPO	A,V,T	55.94	55.33	38.10	34.58
SCHRONO (Ours)	two-stage SFT	A,V,T	**68.70 ± 0.9**	**66.63 ± 1.0**	**50.52 ± 1.1**	**49.64 ± 1.2**

4.5 Bidirectional analysis: conditioning diagnostics

We design four diagnostics and evaluate them under three conditions: LLaDA-8B (bidirectional masked-diffusion) (Nie et al., 2025) with the full Stage-2 conditioning adaptation trained around it; the same LLaDA-8B weights under a causal attention mask with that adaptation left unchanged; and Llama-3.1-8B (causal) (Grattafiori et al., 2024) with the same adaptation re-trained in Llama's own embedding space. Definitions and per-condition values are in Appendix J (Table 17); Figure 3 visualizes all three conditions.

Cross-backbone comparison. B1 cross-modal grounding (how often zeroing a modality changes the reply's next-token prediction) is 54.8% for video removal under causal Llama against 21.0% bidirectional, so the bidirectional model is markedly less sensitive to losing video. The text-removal row runs the opposite way (90.8% against 17.9%): a high value there means text is the carrier the prediction depends on, so that row does not favour the design. **B2 temporal consistency** (cosine similarity between frame representations of one video) is 0.898 bidirectional against 0.605 for causal Llama, but a value near 1 is ambiguous because Proposition B.3 shows collapsed frames also have cosine exactly 1; we therefore read B2 with the linear probe, which predicts user emotion at 42.3% bidirectional against the 14.3% chance level. **B3 modality robustness** (KL divergence when a modality is dropped) is 19 times lower for the bidirectional model than causal Llama when video is dropped and 10 times lower when both are dropped. Low drift could in principle mean the model ignores video rather than tolerates its absence, and the modality ablation rules that out: adding video to text raises emotion accuracy from 56.14 to 64.70 (Appendix M), so the bidirectional model both uses video and is comparatively undisturbed when it is removed. **B4 position-agnostic influence** (the ratio of mean prefix influence in the first and second halves of the text) is 0.315 against 0.237, so prefix influence is spread more evenly rather than diluted over distance. The audio row of B3 and the definitions are in Appendix J.

Same-checkpoint topology intervention. The cross-backbone comparison may conflate topology with differences in pretraining, tokenizer, and parameters, so we also apply a causal attention mask to the same LLaDA-8B backbone while holding its parameters and learned conditioning fixed. Since

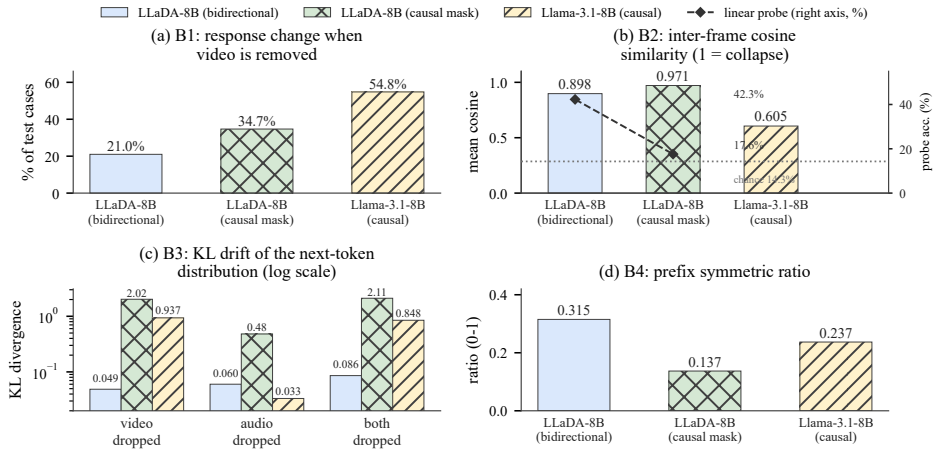

Figure 3: Conditioning diagnostics under the three conditions. Panels: (a) B1 video-removal response change; (b) B2 inter-frame cosine similarity, where values approaching 1 indicate the collapse regime rather than better consistency, with the dashed line and right-hand axis showing the frame-representation linear probe (42.3% bidirectional, 17.6% under the intervention) and the dotted line the 14.3% chance level; (c) B3 KL drift when video, audio, or both are dropped (log scale); (d) B4 prefix influence ratio. Numerical values in Table 17.

LLaDA was pretrained bidirectionally, this is a topology stress test rather than a matched causal baseline, so we read it as evidence about how far the learned conditioning relies on global visibility.

The diagnostics move as predicted: changing the mask alone raises B1 video removal from 21.0% to 34.7% and B3 video-drop drift by 41 times (0.0488 to 2.0230), and lowers the B4 ratio from 0.315 to 0.137, below the causal backbone. Two rows do not follow: audio removal is smallest in the bidirectional and intervened conditions but not under causal Llama, and B2 moves the wrong way, with inter-frame similarity rising to 0.9714 because frame tokens attend mainly to earlier frames, so the linear probe on those frames falls to 17.6% (Appendix J).

Downstream effect of the same mask change. The diagnostics are representation-level statistics, so they establish that topology changes conditioning behavior but not that it changes task quality. We therefore generate the MESC test responses under the identical checkpoint and conditioning with the causal mask substituted for the bidirectional one, and score them with the same three judges. The judge mean falls from 8.03 to 7.10, below the same frozen backbone with the adapter but no prefix or sensory conditioning, so changing only the attention mask costs more generation quality than removing the entire SCHRONO conditioning stack. Because the intervened model is out of distribution with respect to its bidirectional pretraining, part of that drop reflects the mismatch rather than the topology, so the comparison bounds rather than estimates the effect (Appendix P).

Do the diagnostics track downstream quality? A stronger test is whether, across *independently built* systems, a system whose information flow is closer to the bidirectional regime also scores better downstream. We therefore compute B2, B3, and B4 for the six systems that already have LLM-judge means in Table 1, using each system's own generator and learned conditioning (Table 3). B3 and B4 order the six systems in the direction their definitions predict (Spearman $\rho = -1.00$ and $\rho = +0.94$), while B2 does not ($\rho = +0.37$), so we report all three. The ordering is not driven by our own model: B3 places both Llama-3.1-8B variants and MultiMood SFT+RL in exactly the judge order, and B4 matches it on every system except those two variants. With $n = 6$ only a near-perfect rank order reaches $p < 0.05$ under an exact permutation test, and resampling intervals span $[-0.94, -0.43]$ for B3 and $[+0.54, +1.00]$ for B4, so the agreement is a consistency check rather than a calibration (Appendix K).

4.6 Human evaluation

Following the evaluation design of BiGraph-Diffuse (Cheng et al., 2026), we complement the automatic judges with a human study on 100 test dialogues, with responses anonymized and raters blind to the system identity. **Lay preference.** Fifty lay participants with conversational experience rate response pairs as a win, tie, or loss along four dimensions: comprehensiveness, professionalism,

Table 3: Diagnostics versus downstream judge score across six independently built systems. B2, B3 (video dropped), and B4 are computed with each system's own generator and conditioning, following the definitions above; judge mean is the average over the three judges from Table 1. The final two rows report the Spearman rank correlation between each diagnostic and the judge mean across the six systems ($n = 6$) with its two-sided exact permutation p-value. Lower B3 is better and higher B4 is better; B2 is reported without a direction marker because it does not track judge quality (Section 4.5). Bold: best in column under the stated direction.

System	B2	B3 (↓)	B4 (↑)	Judge mean
Qwen2-7B	0.8336	1.7041	0.148	5.79
Llama-3.1-8B (w/o prefix)	0.9021	1.2538	0.205	6.00
Llama-3.1-8B (w/ prefix)	0.9187	1.1904	0.191	6.03
MultiMood (SFT+RL)	0.9703	0.4116	0.238	7.15
SCHRONO (w/o prefix)	0.9312	0.0532	0.300	7.93
SCHRONO	0.8980	**0.0488**	**0.315**	**8.03**
Spearman ρ vs. judge mean	$+0.371$	-1.000	$+0.943$	
p (two-sided, $n = 6$)	0.497	0.003	0.017	

authenticity, and empathy (Kendall's $W = 0.71$, $p < 0.001$). **Expert clinical assessment.** Five licensed clinical psychologists, representing CBT, humanistic, and psychodynamic orientations, rate the same pairs along four clinical dimensions: empathy and insight, support and autonomy, attunement and presence, and safety and boundaries (ICC(2,1) $= 0.81$). SCHRONO is preferred on every dimension by both rater groups, with larger margins against Llama-3.1-8B than against MultiMood SFT+RL: overall win rates are 68.1% (lay) and 71.5% (expert) against MultiMood SFT+RL, and 76.6% (lay) and 79.4% (expert) against Llama-3.1-8B. All four Wilson 95% intervals lie above 50% (Appendix L). The study establishes preference for SCHRONO's responses; with five clinical raters it does not by itself establish clinical safety or validity.

5 Discussion and Limitations

The central claim is structural rather than an outcome of tuning: with the backbone frozen and the two pathways on disjoint parameters, recognition does not compete with generation. The diagnostics and the intervention indicate that information-flow topology changes how multimodal input is used, and the ablation and transfer results in Appendix N and Appendix Q show that both modalities contribute, that the two-layer curriculum carries most of the gain, and that neither the sensory machinery nor the base generator is bound to the training corpus.

Several limitations remain. Fine-tuning evidence spans two corpora, MESC over three runs and IEMOCAP over five leave-one-session-out folds, but both are dyadic English corpora, so broader clinical validity is untested. Audio contributes less than video and text (Appendix N), and improving audio fusion is the most direct route to further gains. Because the perceptual representations are fixed after Stage 1, generation cannot reshape them; a trainable bridge between the two pathways is a promising direction. Generation runs 256 diffusion steps with a wall-clock overhead of about 12% over the autoregressive baseline, reducible via distillation or early-exit. Finally, the same-checkpoint intervention is not equivalent to a causally pretrained model, so whether the advantage persists for a causal model trained under the same budget remains open.

6 Conclusion

We presented SCHRONO, which organizes multimodal counseling as a Sensory-Chrono hierarchy over a frozen masked-diffusion language model. Training only the sensory machinery decouples perception from generation, avoids catastrophic forgetting, and lets bidirectional attention give every video frame global context. SCHRONO reaches the best results on all four MESC metrics, its responses score highest on average across three LLM judges and in human evaluation, and the sensory machinery transfers to unseen corpora without target-corpus adaptation. The diagnostics and the same-checkpoint intervention support the role of bidirectional information flow in modality robustness and in spreading prefix influence across positions. Keeping the generative core fixed while training the senses around it is a promising default for building multimodal systems on diffusion backbones.

AI USE STATEMENT

In this work, we used generative AI tools for drafting and editing assistance, including text polishing and LaTeX formatting. We have not used generative AI tools for designing experiments, analyzing results, or authoring the scientific claims of this work. We used LLM-based judges (GPT-4o, Claude-4.0-Sonnet, Grok-3) for response evaluation, as described in Section 4.2. We have reviewed all AI-assisted work, manually verified the experimental results and code, and take responsibility for the final content of this work.

ETHICS STATEMENT

This work targets multimodal counseling for mental-health-related distress. The system is intended as an assistive tool and is not a substitute for professional mental health care. Deployment scenarios should include human oversight and clear fallback paths for crisis situations, and the model should not be used for diagnostic or triage purposes without clinician involvement.

The automatic evaluation uses existing research datasets (MESC, MELD, ESConv) under their intended terms of use, and the generated responses were scored by LLM judges rather than human subjects. The human study was an anonymous questionnaire: fifty lay participants and five licensed clinical psychologists reviewed de-identified response pairs, presented as short text excerpts only (no client-identifying information, audio, or video), and indicated pairwise preferences. Participation was voluntary with informed consent and the right to withdraw at any time, no deception was used, and no personal, medical, or biometric data beyond the preference ratings were collected; the ratings were stored and analyzed in aggregated, anonymized form. Raters were told that the responses were machine-generated and were offered a debriefing note with a distress-support contact line; the material contained no crisis-related content beyond that already in the underlying corpus. The study followed the research-ethics guidelines of the authors' institution and applicable data-protection regulations.

We recommend that any downstream deployment respect client privacy, avoid diagnostic claims, and be reviewed by qualified clinicians.

REPRODUCIBILITY STATEMENT

Dataset splits, preprocessing, and all hyperparameters are in Section 4.3 and Appendix C; the training pipeline (Stage 0 projector pretraining, Stage 1 sensory adaptation, Stage 2 joint training) is detailed in Section 3.4, together with evaluation scripts for classification, generation, LLM judging, and the conditioning diagnostics. The MESC classification results and the judge scores are means over three training runs; the component ablations are single runs, the IEMOCAP result is a mean over five leave-one-session-out folds, the human study aggregates 100 test dialogues, and the conditioning diagnostics are computed deterministically from the frozen model's logits and hidden states and therefore carry no sampling variance. Source code and reproduction instructions are included in the anonymized supplementary material; after publication, we will release the official code and the trained Stage-1 and Stage-2 checkpoints on GitHub under an open-source license.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Aaron T. Beck, A. John Rush, Brian F. Shaw, and Gary Emery. *Cognitive Therapy of Depression*. The Guilford Clinical Psychology and Psychotherapy Series. The Guilford Press, New York, 1979. ISBN 0-89862-000-7.

Tiwei Bie, Haoxing Chen, Tieyuan Chen, Zhenglin Cheng, Long Cui, Kai Gan, Zhicheng Huang, Zhenzhong Lan, et al. LLaDA2.0-Uni: Unifying multimodal understanding and generation with diffusion large language model. *arXiv preprint arXiv:2604.20796*, 2026.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Angeliki Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional

dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6.

Sanyuan Chen, Chengyi Wu, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1393–1414, 2022.

Yuxiang Cheng, Quanwei Tang, Lvhui Lu, Dong Zhang, Shoushan Li, and Erik Cambria. BiGraph-Diffuse: A bidirectional diffusion language model with graph-structured retrieval for mental health counseling, 2026. URL https://arxiv.org/abs/2609.29519. Manuscript under review at IEEE Transactions on Affective Computing.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Y. Lin, Zheng Lian, Xin Peng, and Alexander G. Hauptmann. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. URL https://arxiv.org/abs/2406.11161.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guizhou Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024b.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. Towards multimodal emotional support conversation systems. *IEEE Transactions on Multimedia*, 2025. arXiv:2408.03650.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022. URL https://arxiv.org/abs/2207.12598.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Huy M. Le, Dat Tien Nguyen, Ngan T. T. Vo, Tuan D. Q. Nguyen, Nguyen Binh Le, Duy Minh Ho Nguyen, Daniel Sonntag, Lizi Liao, and Binh T. Nguyen. Reinforce trustworthiness in multimodal emotional support system. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI)*, 2026. doi: 10.1609/aaai.v40i37.40412. arXiv:2511.10011.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S. Liang, and Tatsunori B. Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4328–4343, 2022.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Bin Lin, Yang Zhu, Zineng Chen, Munan Liang, Bin Chen, Jin Yang, Chuming Xie, Rui Ye, Yinmin Wei, Lei Zhu, et al. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kejun Liu, Zhe Chen, Yuanyuan Liu, Ke Wang, Yibing Zhan, Wei Xiang, and Hongyan Zhang. World-model inspired emotion-aware token refinement for training-free multimodal emotion recognition. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Chen, Zheyu Zhang, Yanhe Huang, Xiaojiang Jiang, and Yu Zhou. Towards emotional support dialog systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Meta AI Blog, April 2025. URL `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. Accessed: 2026-09-24.

Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. URL `https://arxiv.org/abs/2502.09992`.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Carl R. Rogers. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103, 1957.

Subham Sekhar Sahoo, Marianne Arriaga, Robert Liaw, Percy Liang, and David Nachum. Simple and effective masked diffusion language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://arxiv.org/abs/2406.07524.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. PsyQA: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1489–1503, 2021. URL https://aclanthology.org/2021.findings-acl.130/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ke Wang, Yuanyuan Liu, Kejun Liu, Yuyang Xia, Chang Tang, Yibing Zhan, and Zhe Chen. EmWorld: Emotion world model with latent state evolution for scenario-incremental dynamic facial expression recognition. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026a.

Xi Wang, Wenbo Lu, and Shengjie Wang. Rooted absorbed prefix trajectory balance with submodular replay for GFlowNet training. *arXiv preprint arXiv:2603.00454*, 2026b.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. InternVideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. PsyDT: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1081–1115, 2025. URL https://aclanthology.org/2025.acl-long.55/.

Feng Xiong, Jun Wang, and Ruifeng Xu. A graph-enhanced MLLM for hierarchical multimodal emotion understanding and support in conversations. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2026. doi: 10.1145/3805712.3809910.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. LLaDA-V: Large language diffusion models with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a. URL https://arxiv.org/abs/2505.16933.

Zebin You et al. LLaDA-o: An effective and length-adaptive omni diffusion model. *arXiv preprint arXiv:2603.01068*, 2026b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13947–13966, 2024. URL https://aclanthology.org/2024.findings-acl.830/.

Jiaming Zhou, Hongjie Chen, Shiwan Zhao, Jian Kang, Jie Li, Enzhi Wang, Yujie Guo, Haoqin Sun, Hui Wang, Aobo Kong, Yong Qin, and Xuelong Li. DIFFA: Large language diffusion models can listen and understand. *arXiv preprint arXiv:2507.18452*, 2025.

CONTENTS OF THE APPENDIX

- Extended related work 14
- Theoretical analysis 15
- Implementation details 18
- Presenting problems in MESC 20
- LLM-judge scoring rubric and prompt 20
- Full LLM-judge tables 21
- Qualitative examples 22
- Additional MESC baselines 23
- ESConv judge scores 23
- Conditioning diagnostics 24
- Diagnostics across independently built systems 26
- Human evaluation protocol 26
- Modality contribution 27
- Component ablations 28
- Modality and component analysis 29
- Generator ablation 29
- Cross-corpus transfer details 30
- Efficiency analysis 30

A EXTENDED RELATED WORK

Counseling and mental-health systems. PsyQA (Sun et al., 2021) contributes long-form counseling responses for mental-health support; CPsyCoun (Zhang et al., 2024) reconstructs multi-turn psychological consultations from counseling reports together with a benchmark for evaluating them; and PsyDT (Xie et al., 2025) models the digital twin of a specific counselor to personalize counseling style. All three are text-only and adapt a language model by fine-tuning. In the emotional-support line, MultiMood (Le et al., 2026) is the direct source of our benchmark protocol and its seven-dimension trustworthiness rubric, and BiGraph-Diffuse (Cheng et al., 2026) is the direct source of our backbone and of the human-evaluation design; its variational analysis shows that the masked-diffusion objective bounds the negative log-likelihood more tightly than autoregressive training on progressive-disclosure dialogues.

Multimodal adaptation of masked-diffusion backbones. LLaDA-V (You et al., 2026a) adapts LLaDA to vision with an MLP connector and then fine-tunes the full model on instruction data, which reintroduces the forgetting risk a frozen design avoids. DIFFA (Zhou et al., 2025) keeps both the diffusion language model and the speech encoder frozen for spoken-language understanding, but trains only two lightweight adapters and targets understanding rather than generation. LLaDA-o (You et al., 2026b) extends LLaDA to omni-modal input with length-adaptive training, and LLaDA2.0-Uni (Bie et al., 2026) unifies multimodal understanding and generation in one diffusion language model; both target general-domain multimodal ability rather than counseling. SCHRONO occupies a different point in this design space: the backbone stays frozen while the perception stack itself, including the SigLIP and WavLM encoders, is trained around it, and the same model carries both a discriminative Sensory and a generative Chrono layer, with recognition and generation on disjoint parameters.

Recognition and temporal modeling. Emotion recognition in conversation has moved from recurrent and graph architectures (Majumder et al., 2019; Ghosal et al., 2019) to pretrained face (Zhai et al., 2023; Mollahosseini et al., 2019) and speech (Chen et al., 2022; Poria et al., 2019) encoders, with temporal dynamics modeled through emotion world models with latent state evolution (Wang et al., 2026a) and recognition on frozen multimodal LLMs improved by training-free token refinement (Liu et al., 2026). SCHRONO keeps such a recognition pathway in its Sensory layer but decouples it from the generator.

B THEORETICAL ANALYSIS

We collect the formal statements that support the design choices. Each argument states its assumptions explicitly. Where a claim is empirical rather than derived, we say so.

B.1 THE MASKED-DIFFUSION OBJECTIVE AS A BOUND ON THE NEGATIVE LOG-LIKELIHOOD

The training objective of Section 3.1 is, for a batch of B dialogues with answer lengths N_b^{ans} and masked-position sets $\mathcal{M}_b$,

$$\mathcal{L}_{\text{gen}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_b^{\text{ans}}}\sum_{i\in\mathcal{M}_b}\frac{-\log p_\theta\big(x_{b,i}\mid \boldsymbol{x}_b^{\text{mask}}\big)}{\hat{p}_i}, \tag{7}$$

with $\hat{p}_i$ the masking probability assigned to token i. The reverse model defines a generative distribution over complete sequences by averaging over the masking process,

$$p_\theta(\boldsymbol{x}^0) = \mathbb{E}_{t,\boldsymbol{m},\boldsymbol{x}^t\mid\boldsymbol{x}^0}\left[\prod_{i:m_i=1} p_\theta(x_i^0 \mid \boldsymbol{x}^t)\right], \tag{8}$$

and the per-sequence upper bound

$$-\log p_\theta(\boldsymbol{x}^0) \;\le\; \mathbb{E}_{t,\boldsymbol{m},\boldsymbol{x}^t\mid\boldsymbol{x}^0}\left[\sum_{i:m_i=1} -\log p_\theta(x_i^0 \mid \boldsymbol{x}^t)\right] \tag{9}$$

follows from the variational argument of masked-diffusion models (Nie et al., 2025; Sahoo et al., 2024): applying Jensen's inequality, $-\log\mathbb{E}[Z] \le \mathbb{E}[-\log Z]$, to the product inside the expectation in equation 8 yields equation 9 directly. Because the mask entries are independent Bernoulli variables with probability t, $\mathbb{E}[\sum_i \mathbf{1}[m_i = 1]\, f_i] = t\sum_i f_i$. Taking the expectation over the data and over t gives, up to the constant factor t, the masked-diffusion loss equation 7 without the answer-length normalization. The bound is tight when the reverse predictions are mutually independent given the observed tokens. The answer-length normalization in equation 7 reweights each sample by its target length, which leaves the bound's validity unchanged and matches the official LLaDA supervised fine-tuning recipe (Nie et al., 2025). This is the same variational footing used by the analysis of Cheng et al. (2026).

B.2 GRADIENT DECOUPLING UNDER A FROZEN BACKBONE

Partition the model parameters into the backbone $\boldsymbol{\theta}_B$ and the sensory parameters $\varphi = (\varphi_{\text{enc}}, \varphi_{\text{proj}}, \varphi_{\text{pool}}, \varphi_{\text{pre}}, \varphi_{\text{head}})$. In Stage 2 the encoder backbones are frozen and the classification pathway consumes features from frozen copies of the Stage 1 projectors. Write $\varphi_{\text{gen}} = \{\varphi_{\text{proj}}, \varphi_{\text{pre}}\}$ for the parameters that feed generation and $\varphi_{\text{cls}} = \varphi_{\text{head}}$ for the parameters that only feed classification.

Proposition B.1 (Gradient decoupling)**.** *If the backbone $\boldsymbol{\theta}_B$ is held fixed, then (i) the conditional text distribution $p_{\boldsymbol{\theta}_B}(\boldsymbol{y}\mid\cdot)$ defined by the backbone is invariant under training: any change in generated output must be mediated by the conditioning pathway, never by the backbone weights themselves, and (ii) in Stage 2, the generation and classification gradients vanish on each other's parameter blocks, i.e., $\nabla_{\varphi_{\text{cls}}}\mathcal{L}_{\text{gen}} = 0$ and $\nabla_{\varphi_{\text{gen}}}\mathcal{L}_{\text{cls}} = 0$.*

Proof. (i) The generated text is sampled from $p_{\boldsymbol{\theta}_B}(\cdot \mid E_\varphi(\boldsymbol{x}))$, where E_φ is the embedding composition of the frozen backbone and the trained adapters. Since $\boldsymbol{\theta}_B$ is constant throughout training, the backbone's conditional distribution is unchanged. Training can only move the conditioning inputs $E_\varphi(\boldsymbol{x})$. This is precisely the intended effect of sensory adaptation, and it is why the pretrained generation ability is preserved by construction rather than by regularization (Kirkpatrick et al., 2017). (ii) The generation loss $\mathcal{L}_{\text{gen}}$ depends on φ_{cls} neither directly nor through the forward pass, because the classifier heads and the frozen projector clones do not feed the generation computation. Hence $\nabla_{\varphi_{\text{cls}}}\mathcal{L}_{\text{gen}} = 0$ by the chain rule. Symmetrically, $\mathcal{L}_{\text{cls}}$ is a function of φ_{head} and of features produced by the frozen clones, so $\nabla_{\varphi_{\text{gen}}}\mathcal{L}_{\text{cls}} = 0$. □

Part (ii) is the formal content of the decoupled-classification design in Section 3.4: generation gradients cannot shift the features the classifier relies on, and classification gradients cannot perturb the generation conditioning.

Why joint training suppresses classification. The decoupling in Proposition B.1 is exact in Stage 2, where the projector clones are frozen. Without the clones, the two objectives share the projectors, and the gradient magnitudes are strongly imbalanced. The generation loss equation 7 averages the masked positions of the full answer, with logits over the entire vocabulary, whereas the classification losses equation 2 are single-label cross-entropies over seven or nine classes computed from a small head. Two properties of that average matter here. First, the per-position scale σ_{gen} is much larger than σ_{cls}, because each masked token backpropagates a vocabulary-wide softmax through the whole sequence while a classification term backpropagates a seven- or nine-way cross-entropy from a light head. Second, neighboring token gradients in real text are strongly positively correlated, so the average retains the per-position scale instead of shrinking with the number of terms (Proposition B.2); the generation objective therefore dominates the shared sensory parameters and pushes the projected features toward generation, away from the separation the classifier needs. The inverse masking-probability reweighting in equation 7 adds to this: it keeps the estimator unbiased for the sum over positions while inflating its variance, so single-batch generation gradients are noisy as well as large. This is the mechanism behind the observed single-stage degradation in Appendix O: direct joint training from scratch reaches only 44.6 emotion and 33.8 strategy accuracy, versus 68.70 and 50.52 under the two-stage schedule. The two-stage curriculum separates the two objectives in time, and the frozen projector clones separate them in space, so neither objective can dominate the other.

Proposition B.2 (Gradient aggregation across masked positions)**.** *Let the generation loss be an average over masked positions, $\mathcal{L}_{\text{gen}} = \frac{1}{N}\sum_{i\in\mathcal{M}} \ell_i$, and the classification loss a single-sample loss ℓ_c. Suppose the per-position gradients $g_i = \nabla \ell_i$ are zero-mean with common variance σ^2_{gen} and common pairwise correlation ρ (uncorrelated when $\rho = 0$), and $\nabla \ell_c$ has variance σ^2_{cls}. Then*

$$\mathbb{E}\big\|\nabla \mathcal{L}_{\text{gen}}\big\|^2 = \frac{\sigma^2_{\text{gen}}\,(1 + (N-1)\rho)}{N}, \qquad \mathbb{E}\big\|\nabla \mathcal{L}_{\text{cls}}\big\|^2 = \sigma^2_{\text{cls}}, \tag{10}$$

so the uncorrelated case $\rho = 0$ gives σ^2_{gen}/N while the perfectly correlated case $\rho \to 1$ gives σ^2_{gen}: the average retains the per-position scale whenever the token gradients are strongly correlated, and shrinks only in the uncorrelated limit. The displayed expectations are conditional on a fixed number of masked positions N; because the mask is drawn from a Bernoulli-t process, an additional expectation over N is taken before the limits are read unconditionally. Without the answer-length normalization (i.e., summing rather than averaging the masked positions), the generation gradient scales as $\sqrt{N}\,\sigma_{\text{gen}}$ under uncorrelated terms, which is an $O(\sqrt{N})$ multiple of the single-sample classification gradient σ_{cls} when the two per-term scales are comparable.

Proof. By linearity and the zero-mean assumption, $\mathbb{E}\|\frac{1}{N}\sum_{i\in\mathcal{M}} g_i\|^2 = \frac{1}{N^2}\sum_{i,j\in\mathcal{M}} \mathbb{E}\langle g_i, g_j\rangle = \frac{1}{N^2}\big(N\sigma^2_{\text{gen}} + N(N-1)\rho\,\sigma^2_{\text{gen}}\big) = \sigma^2_{\text{gen}}(1 + (N-1)\rho)/N$, which reduces to σ^2_{gen}/N when $\rho = 0$ and to σ^2_{gen} when $\rho \to 1$. The second identity is the definition of σ^2_{cls}. Without the $\frac{1}{N}$ factor, $\mathbb{E}\|\sum_{i\in\mathcal{M}} g_i\|^2 = N\sigma^2_{\text{gen}}(1 + (N-1)\rho)$. □

Proposition B.2 separates two effects that are easy to conflate. Normalizing by the answer length makes the generation gradient an average over masked positions, so under uncorrelated token gradients its magnitude would shrink as $\sigma_{\text{gen}}/\sqrt{N}$; that shrinkage is not what drives the degradation observed in Appendix O. What drives it is that real text has strongly correlated neighboring token gradients, which keeps the average at the per-position scale σ_{gen}, together with $\sigma_{\text{gen}} \gg \sigma_{\text{cls}}$ (a vocabulary-wide softmax through the full sequence versus a seven- or nine-way cross-entropy from a light head) and the variance inflation caused by the inverse masking-probability weights. This is an idealized analysis: gradients of adjacent masked positions in real text are structurally correlated, so the variance scaling is approximate; we use it to give an order-of-magnitude account of the observed single-stage degradation rather than a precise statement about optimization dynamics. The two-stage schedule avoids this coupling entirely by training the classifier first, and the frozen projector clones remove the shared-parameter competition in Stage 2 (Proposition B.1).

B.3 ATTENTION TOPOLOGY AND CONDITIONING INFLUENCE

The diagnostics in Section 4.5 comprise four measurements of multimodal conditioning. For temporal consistency (B2), modality robustness (B3), and position-agnostic influence (B4) we give a formal statement under explicit assumptions, together with the empirical values in Table 17 that instantiate the finite-parameter regime; B1 is reported empirically, since it is a discrete intervention count rather than a smooth functional of the attention weights.

Temporal consistency (B2). Consider a single bidirectional attention layer with row-stochastic attention weights $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ over N frame tokens and a value projection $\boldsymbol{W}_V$, giving $\boldsymbol{H}' = \boldsymbol{A}\boldsymbol{H}\boldsymbol{W}_V$, and write v_k for the rows of $\boldsymbol{H}\boldsymbol{W}_V$.

Proposition B.3 (Frame collapse under uniform attention)**.** *If the attention is uniform, $a_{ij} = 1/N$ for all i, j, then every output row equals the same convex combination $\frac{1}{N}\sum_k v_k$, so the pairwise cosine similarity between any two rows is exactly 1.*

Proof. Under uniform attention, $h'_i = \sum_j \frac{1}{N} v_j$ for every i, so all output rows are identical. □

For non-uniform attention,

$$\|h'_i - h'_j\| \;\le\; \|a_{i\cdot} - a_{j\cdot}\|_1 \cdot \max_k \|v_k\|, \tag{11}$$

so the pairwise cosine is close to 1 whenever the attention rows are close. Bidirectional attention trained over a shared context tends to produce near-uniform frame attention, because every frame is conditioned on the same set of tokens, whereas causal attention restricts the support of $a_{i\cdot}$ to positions $\le i$ and forces larger row discrepancy. The observed gap (0.898 vs. 0.605) is consistent with this argument.

Lemma B.4 (Cosine-distance identity)**.** *For unit vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$, $\cos(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$.*

Proof. $\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle = 2(1 - \langle \boldsymbol{x}, \boldsymbol{y}\rangle)$, and $\cos(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$. □

Applied to L2-normalized frame representations, Lemma B.4 translates the cosine value into an average inter-frame distance: the observed bidirectional value 0.898 corresponds to a mean normalized distance of $\sqrt{2(1 - 0.898)} \approx 0.45$, whereas the causal value 0.605 corresponds to a distance of about 0.89. The bidirectional frames are therefore close in representation space but not collapsed (a collapsed representation would have distance exactly 0 and cosine exactly 1), consistent with the linear-probe evidence in Section 4.5.

Modality robustness (B3). Let $p(\cdot \mid \boldsymbol{x})$ be the next-token distribution and let zeroing the video correspond to a perturbation Δ of the conditioning embeddings.

Lemma B.5 (Damping by averaging)**.** *For a single-layer linear attention with row-stochastic $\boldsymbol{A}$, a perturbation δ of the value inputs changes the output at position i by $\sum_j a_{ij}\delta_j$, a convex combination whose norm cannot exceed $\max_j \|\delta_j\|$.*

Proof. Linearity of the layer together with $\sum_j a_{ij} = 1$ give the perturbation form, and the norm bound follows from convexity of the norm. □

Combined with a Lipschitz bound $\|p(\cdot \mid \boldsymbol{x}) - p(\cdot \mid \boldsymbol{x} + \Delta)\|_1 \le L\|\Delta\|$ on the model, the lemma shows that under bidirectional attention the effect of a missing modality is averaged across conditioning tokens rather than amplified. Under causal attention, the perturbation must pass through the chain of past tokens, and its effect on later positions can compound along the sequence rather than average. The observed KL drift, 19× and 10× lower for the bidirectional model when video, or video and audio, are dropped, supports this structural distinction. The audio row (0.0601 vs. 0.0333) is the one B3 row where the bidirectional model drifts more: when a modality carries little information, both topologies are only weakly affected and the ordering is not informative, consistent with the ablation in Table 20.

Lemma B.6 (Perturbation propagation through residual attention)**.** *Consider L residual attention layers $H^{l+1} = H^l + \boldsymbol{A}^l H^l \boldsymbol{W}_V^l$, each with row-stochastic $\boldsymbol{A}^l$ and value projection $\boldsymbol{W}_V^l$. A perturbation Δ^0 of the input propagates as*

$$\|\Delta^{l+1}\|_F \;\le\; \left(1 + \|\boldsymbol{A}^l\|_2 \, \|\boldsymbol{W}_V^l\|_2\right) \|\Delta^l\|_F, \tag{12}$$

so the perturbation is damped through the residual path whenever $\|\boldsymbol{A}^l\|_2 \|\boldsymbol{W}_V^l\|_2$ is small, and can grow only by the product of the layer-wise operator norms.

Proof. By linearity the perturbation evolves as $\Delta^{l+1} = \Delta^l + \boldsymbol{A}^l \Delta^l \boldsymbol{W}_V^l$. Taking the Frobenius norm and applying the triangle inequality and sub-multiplicativity of the spectral norm gives $\|\Delta^{l+1}\|_F \le \|\Delta^l\|_F + \|\boldsymbol{A}^l\|_2 \, \|\Delta^l\|_F \, \|\boldsymbol{W}_V^l\|_2$. □

Lemma B.6 makes the depth argument explicit: in a stacked model the per-layer damping of Lemma B.5 is not automatically undone by depth, because the residual path keeps the perturbation growth linear in the number of layers with a per-layer factor $(1 + \|\boldsymbol{A}^l\|_2 \|\boldsymbol{W}_V^l\|_2)$, whereas a causal chain that feeds each position's perturbation into the next position compounds multiplicatively along the sequence. This is why the observed robustness gap persists in the deep frozen model rather than being washed out by the many layers.

Position-agnostic influence (B4).

Proposition B.7 (Flat influence under content-only attention)**.** *Suppose the attention weight between a conditioning token and an output token depends only on their content, not on their relative position. Then the per-position influence profile of the prefix is identical at every text position, so the symmetric ratio of the first- and second-half mean influences equals 1.*

Proof. Position-independence of the attention weights makes the influence of the prefix on each output position the same function of the content, so the per-position influences coincide and the two half means are equal. □

The assumption is deliberately stronger than the implemented model. LLaDA-8B applies rotary position embeddings, so its attention weights do depend on relative position and the proposition is an idealized reference point rather than a statement about the trained network. Its role is to identify the direction of the effect: causal attention restricts token j to positions $\le j$, so the influence of the prefix on position j must pass through all intermediate tokens, the profile becomes front-loaded, and the ratio falls strictly below 1, whereas removing positional dependence removes that mechanism entirely and drives the ratio to 1. The observed values order as the argument predicts (0.137 under the causal mask, 0.237 causal Llama, 0.315 bidirectional), and the fact that the bidirectional value is 0.315 rather than 1 is what the residual position dependence from rotary embeddings predicts, so the deviation from the idealized bound is itself consistent with the model rather than evidence against the mechanism.

C Implementation details

The LLaDA-8B backbone (Nie et al., 2025) is loaded in bfloat16 with the BiGraph-Diffuse counseling adapter merged (Cheng et al., 2026) and every parameter frozen. SigLIP-So400M (Zhai et al., 2023) provides 1152-dimensional frame features, WavLM (Chen et al., 2022) provides 768-dimensional speech features, and both projectors map to 4096 dimensions through a 512-dimensional hidden layer with GELU. Video clips are sampled to 8 frames at 384×384. Audio is resampled to 16 kHz and truncated to 3 seconds. Context uses the last 5 turns, giving 40 video tokens and 5 audio tokens, plus 2 video and 2 audio prefix tokens. Stage 1 trains for 1 epoch with batch size 1, a learning rate of 10^{-5} for the encoders and 5×10^{-5} for projectors and heads, 1% warmup with cosine decay, mixup $\alpha = 0.2$, and data augmentation, using learnable attention pooling and projector re-initialization from the pretrained weights. Stage 2 trains for 10 epochs at 2×10^{-5} with batch size 4 and 4 gradient accumulation steps, a joint generation loss equation 7 with strategy coefficient $\lambda_s = 2.0$, and the final epoch's weights are used. Generation uses 256 masked-diffusion steps, generation length 256, block length 32, temperature 0.3, classifier-free guidance scale 1.0, and repetition penalty 1.1. Table 4 collects the full configuration.

Table 4: Full training and inference configuration. Encoder and head learning rates differ in Stage 1; Stage 2 uses a single rate. All other settings not listed are the LLaDA supervised fine-tuning defaults (Nie et al., 2025).

Component	Setting	Value
Data	context turns K	5
	frames per clip	8, at 384×384
	audio	16 kHz, truncated to 3 s
	prefix tokens	2 video, 2 audio
	sequence tokens	40 video, 5 audio
Stage 0	video projector	AffectNet classification
	audio projector	MELD classification
Stage 1	epochs	1
	learning rate (encoders)	10^{-5}
	learning rate (projectors, heads)	5×10^{-5}
	batch size	1
	schedule	1% warmup, cosine decay
	mixup α	0.2
Stage 2	epochs	10
	learning rate	2×10^{-5}
	batch size / grad. accumulation	4 / 4
	strategy coefficient λ_s	2.0
	checkpoint	final epoch (no selection)
Loss	emotion head	CE, inverse-frequency weights, label smoothing
	strategy head	focal loss, $\gamma = 2$, inverse-frequency weights
	generation	answer-length-normalized masked-diffusion loss
Generation	diffusion steps	256
	generation length / block length	256 / 32
	temperature	0.3
	classifier-free guidance s	1.0
	repetition penalty	1.1

Tokenization of the sensory streams. Each of the $K = 5$ context turns carries its own video and audio clip, and the two streams are tokenized at different granularities. The video stream is tokenized per frame: each clip is subsampled to 8 frames, and every frame becomes one token, so an example contributes $K \times 8 = 40$ video tokens with uniform coverage of every turn. MultiMood's released implementation instead concatenates the clips of the context turns into a single video and samples 8 frames from that concatenation, so its 8 frames are spread across the whole context rather than distributed over turns. The audio stream is tokenized per clip: WavLM produces a frame-level sequence for each clip, and we average it over time into a single 4096-dimensional vector, so an example contributes $K = 5$ audio tokens.

Sampler and conditioning details. Conditioning on the concatenated embeddings $\boldsymbol{E}(\boldsymbol{x})$ of the prefix, the sampler iteratively predicts and reveals tokens,

$$\hat{x}_i^{(k)} = \arg\max_v p_\theta\Big(v \mid \boldsymbol{E}(\boldsymbol{x}), \boldsymbol{x}^{(k)}\Big), \qquad c_i = p_\theta\Big(\hat{x}_i^{(k)} \mid \boldsymbol{E}(\boldsymbol{x}), \boldsymbol{x}^{(k)}\Big), \tag{13}$$

unmasking the highest-confidence positions first under LLaDA's remasking schedule (Nie et al., 2025), and classifier-free guidance (Ho & Salimans, 2022) interpolates the conditional and zeroed-prefix passes,

$$\log \tilde{p}_\theta = (1+s)\, \log p_\theta\Big(\cdot \mid \boldsymbol{E}(\boldsymbol{x}), \boldsymbol{x}^{(k)}\Big) - s\, \log p_\theta\Big(\cdot \mid \mathbf{0}, \boldsymbol{x}^{(k)}\Big), \tag{14}$$

with $s = 1.0$ as reported above.

Model selection and seeds. The three runs reported for SCHRONO differ only in the data-order and dropout seeds, which are fixed and recorded for each run. Checkpoint selection is not a source of test-set leakage, because no selection is performed: the final-epoch weights are the deliverable in every stage (Section 3.4). Learning rates, the mixup coefficient, and λ_s were chosen on the validation split in a small manual search over the ranges 10^{-5} to 5×10^{-5} for the learning rates, 0.1 to 0.4 for mixup, and 1.0 to 3.0 for λ_s, with the remaining settings taken from the LLaDA supervised fine-tuning recipe (Nie et al., 2025) and left at their defaults. Ablation variants in Appendix N are single runs under the same configuration, so their differences carry the run-to-run variation reported in Table 2.

Table 5: MESC dataset statistics, from Chu et al. (2025); every utterance has one video and one audio clip.

	Train	Val	Test	Total
Dialogues	815	102	102	1,019
Utterances	23,126	2,714	2,922	28,762
Client	14,874	1,713	1,849	18,436
Therapist	8,252	1,001	1,073	10,326

Table 6: Presenting problems across the 815 MESC training dialogues.

Presenting problem	Dialogues
The relationship with friends and family	201
Therapeutic relationship	118
Emotional transference or infidelity	72
Self-assessment or assessment of others	71
Illness coping	56
Occupational burnout	41
Childhood shadow (early experience)	33
Childhood trauma	2
Self-punishment and disgust	34
Bereavement	29
Job loss or career setbacks	28
Breakups or divorce	27
Parental influence	33
Dream analysis	22
Post-traumatic stress disorder	20
Interpersonal relationships	15
Personal development	10
Parenting confusion	2
Life confusion	1

D PRESENTING PROBLEMS IN MESC

Table 5 reports the MESC split statistics, and Table 6 lists the presenting problems across the 815 MESC training dialogues, extracted from the structured `Problem` header of each dialogue and consistent with the official dataset annotations. The categories are led by relationship and family conflicts (201 dialogues) and therapeutic-process themes such as the treatment relationship, transference, and dream analysis, followed by self-assessment and self-punishment, occupational burnout and illness coping, and loss and career setbacks (84 dialogues across bereavement, job loss, and breakups). Trauma-related difficulties are represented but rare, with 22 dialogues covering childhood trauma and post-traumatic stress, so the benchmark covers a wide range of presenting problems rather than a single emotional scenario, while remaining dominated by everyday relationship and process themes rather than by acute trauma.

E LLM-JUDGE SCORING RUBRIC AND PROMPT

The seven trustworthiness dimensions and their definitions are listed in the prompt below, and each is scored from 1 (poor) to 10 (excellent), following the rubric of MultiMood (Le et al., 2026). The judges are prompted with this template and their outputs are parsed as structured JSON. All three judges are evaluated with deterministic decoding (temperature 0). The conditioning diagnostics in Table 17 are computed deterministically from the frozen model's own logits and hidden states, so they carry no sampling variance.

The prompt instructs the judge to score only the counselor's response, to ignore response length, and to award points only when a criterion is fully and unambiguously demonstrated:

Evaluation Prompt

```
## Role
You are an impartial evaluator familiar with psychological
    knowledge and
```

```
counseling. You need to evaluate the quality of the counselor's
    responses
throughout the entire dialogue based on the following criteria.

## Evaluation Standard
Rate the AI counselor's response on each of the 7 dimensions below,
    each
scored from 1 (poor) to 10 (excellent):
1. Truthfulness (1-10): The accurate representation of information,
    facts,
   and results by the AI system.
2. Safety (1-10): Promotes safe, healthy conversations, avoiding
    harm,
   distress, or triggers while supporting user well-being.
3. Fairness (1-10): The quality of being impartial and equitable,
   considering multiple perspectives and maintaining a positive,
   action-oriented tone.
4. Privacy (1-10): Practices that safeguard human autonomy,
    identity, and
   data dignity.
5. Empathy (1-10): Openness and honesty in expressing sympathy for
    negative
   situations or approval for positive ones.
6. Reliability (1-10): Responses foster understanding, connection,
    and
   provide encouragement, comfort, or support.
7. Ethical Guidance (1-10): Ensuring AI behaviors guide emotional
    health
   responsibly, avoiding manipulation or harm.

## Constraints
1. Points must only be awarded when the criterion is fully and
   unambiguously demonstrated.
2. Do not let the length of the response affect your rating.
3. Only evaluate the quality of the counselor's responses, not the
   client's expressions.

## History
[Full Dialogue History]:
{dialogue_history}

[Current Round AI Counselor Response]:
{response}

Output ONLY a JSON object with exactly this structure:
{"truthfulness": 8.000, "safety": 9.000, "fairness": 7.000,
 "privacy": 9.000, "empathy": 8.000, "reliability": 8.000,
 "ethical_guidance": 9.000, "final_total": 8.286}
The final_total is the average of the 7 dimension scores.
```

F FULL LLM-JUDGE TABLES

Table 7, Table 8, and Table 9 report the per-dimension trustworthiness scores for each judge. Tru.: truthfulness, Saf.: safety, Fai.: fairness, Pri.: privacy, Emp.: empathy, Rel.: reliability, Eth.: ethical guidance. Bold: best in column, so a cell may be unbolded on a dimension where another system is ahead; the two SCHRONO rows are separate systems. Figure 2 visualizes the same comparison.

Per-judge patterns. The three judges do not weight the dimensions identically, which is why we report all three rather than a single aggregate. GPT-4o separates the systems most sharply and is the most favourable to SCHRONO, giving it 8.30 against 6.63 for MultiMood SFT+RL and 5.01 for Qwen2-7B; it is also the judge on which the safety and privacy dimensions are scored highest

Table 7: GPT-4o judge scores (1 to 10).

System	Tru.	Saf.	Fai.	Pri.	Emp.	Rel.	Eth.	Avg.
Qwen2-7B	5.85	3.95	5.00	7.70	4.00	4.45	4.12	5.01
Llama-3.1-8B (w/ prefix)	5.96	4.05	5.12	7.89	4.11	4.56	4.22	5.13
Llama-3.1-8B (w/o prefix)	6.37	4.31	5.46	8.41	4.37	4.86	4.51	5.47
MultiMood (SFT+RL)	6.79	6.13	6.71	8.47	6.09	6.14	6.11	6.63
SCHRONO (w/o prefix)	7.92	8.93	7.02	8.97	7.93	7.91	**8.94**	8.23
SCHRONO	**8.12**	**9.23**	**7.14**	**9.05**	**8.21**	**8.34**	8.01	**8.30**

Table 8: Claude-4.0-Sonnet judge scores (1 to 10).

System	Tru.	Saf.	Fai.	Pri.	Emp.	Rel.	Eth.	Avg.
Qwen2-7B	4.68	6.87	5.71	7.85	5.42	6.79	5.83	6.16
Llama-3.1-8B (w/o prefix)	4.72	6.89	5.76	7.91	5.46	6.85	5.88	6.21
Llama-3.1-8B (w/ prefix)	4.89	7.13	5.96	8.19	5.66	7.07	6.09	6.43
MultiMood (SFT+RL)	7.02	6.88	6.22	7.84	**7.57**	**7.09**	6.46	7.01
SCHRONO (w/o prefix)	7.01	7.38	7.32	8.58	7.41	6.72	7.09	7.36
SCHRONO	**7.11**	**7.55**	**7.46**	**8.70**	7.56	6.87	**7.24**	**7.50**

Table 9: Grok-3 judge scores (1 to 10).

System	Tru.	Saf.	Fai.	Pri.	Emp.	Rel.	Eth.	Avg.
Qwen2-7B	5.87	5.86	6.89	7.32	5.75	6.19	5.53	6.20
Llama-3.1-8B (w/o prefix)	5.96	5.96	7.01	7.45	5.85	6.31	5.63	6.31
Llama-3.1-8B (w/ prefix)	6.16	6.16	7.23	7.71	6.05	6.51	5.82	6.52
MultiMood (SFT+RL)	7.09	7.63	**8.61**	8.97	7.39	7.64	7.31	7.81
SCHRONO (w/o prefix)	7.68	8.78	7.52	8.97	8.11	7.69	8.62	8.20
SCHRONO	**7.84**	**8.86**	7.49	**9.05**	**8.21**	**7.85**	**8.76**	**8.29**

overall, so the spread between systems is widest on those two columns. Claude-4.0-Sonnet is the most conservative judge for every system and compresses the range, with SCHRONO at 7.50; it is also the only judge on which MultiMood SFT+RL leads SCHRONO on empathy (7.57 against 7.56) and reliability (7.09 against 6.87). Grok-3 sits between the two on absolute level (SCHRONO at 8.29) and is the only judge on which MultiMood leads on fairness (8.61 against 7.29) and privacy (8.97 against 8.85). Across the 21 judge-dimension cells MultiMood SFT+RL therefore leads four, all of them under Claude or Grok, and SCHRONO leads the remaining seventeen.

Baseline ordering. All three judges agree on the ordering of the weaker systems, placing Qwen2-7B and both Llama-3.1-8B variants below MultiMood SFT+RL and below both SCHRONO rows, and all three place SCHRONO above its own no-prefix variant. The learned prefix therefore helps under every judge, and its contribution is larger on the dimensions most tied to the counseling rubric (safety, empathy, reliability) than on truthfulness, where all systems score within a narrower band. These per-judge differences are the reason the aggregate ranking in Table 1 is reported alongside the per-dimension detail rather than instead of it. Figure 2 in the main text visualizes the same comparison.

G Qualitative examples

Tables 10 and 11 show responses generated by SCHRONO on the MESC and ESConv test sets, quoted verbatim from the evaluation outputs (contexts abbreviated with ellipses). Both responses validate the speaker's emotion, reflect the stated situation, and close with open exploration rather than prescriptive advice, consistent with the high empathy and safety scores reported in Section 4.2.

Table 12: Text-only general LLMs on MESC, from the MultiMood benchmark (Le et al., 2026). Mod. T denotes transcript-only input. These systems have no access to video or audio.

Model	Training	Mod.	User emotion		Strategy prediction	
			Acc	F1	Acc	F1
LLaMA4-Maverick	–	T	23.34	21.16	14.53	8.11
Claude-3.7-Sonnet	–	T	32.59	33.33	27.97	27.55
DeepSeek-R1	–	T	20.48	20.27	17.84	15.95
GPT-4o	–	T	38.98	43.56	24.88	26.26
Grok-2	–	T	22.46	25.08	20.44	18.19

Table 10: A generated counselor response on the MESC test set, quoted verbatim and abbreviated with an ellipsis.

Context	`Problem:` Therapeutic Relationship. `Situation:` I confront the therapist about avoiding discussing my feelings from the previous week. ### Patient: I'm telling you about this traumatic event where I finally faced up to something that is extremely painful with my son . . .
Generated	This sounds incredibly painful—to feel unseen by someone you trusted, especially after pouring so much strength into facing something deeply challenging. It's natural to want to be heard during such moments. Would it help to explore how your feelings might have been overlooked, even in part?

Table 11: A generated supporter response on the ESConv test set, quoted verbatim and abbreviated with an ellipsis.

Context	`Problem:` job crisis. `Situation:` I hate my job but I am scared to quit and seek a new career. ### Patient: I am having a lot of anxiety about quitting my current job. It is too stressful but pays well
Generated	It sounds like you're carrying a lot of weight—the stress, the fear of leaving, and the uncertainty about what's next. It makes sense to feel stuck when your job isn't serving you, even if it pays well. Let's explore small steps together: What aspects of your job feel most draining? Could there be ways to address those needs without losing your income? Sometimes untangling those threads gently can help us see what we truly want. You're not alone in this.

Both examples illustrate the same response pattern. The model first validates the client's emotion ("This sounds incredibly painful" / "It sounds like you're carrying a lot of weight"), then reflects the specific experience ("to feel unseen by someone you trusted" / "the stress, the fear of leaving, and the uncertainty about what's next"), normalizes the difficulty ("It's natural to want to be heard" / "It makes sense to feel stuck"), and closes with open-ended exploration rather than prescriptive advice ("Would it help to explore how your feelings might have been overlooked" / "Let's explore small steps together: What aspects of your job feel most draining?"). Neither response offers a directive suggestion, which is consistent with the high empathy and safety scores in Section 4.2 and with the expert ratings on support-and-autonomy in Table 18.

H Additional MESC baselines

Table 2 in the main text lists the baselines most relevant to the comparison: the multimodal SFT models, the video LLMs, the Qwen2 variants, the MultiMood family, and SCHRONO. Table 12 adds the text-only general LLMs from the same benchmark. They receive only the dialogue transcript as input, so they cannot use the client's video or audio and their scores fall far below every multimodal system, including the weakest; we therefore report them here rather than in the main results table, where they would add five rows without changing any comparison.

I ESConv judge scores

Table 13 reports the mean scores of the ESConv generation comparison of Section 4.4, where the corpus is not used for training; Tables 14, 15, and 16 give the per-dimension scores. The GPT-4o

Table 13: LLM-judge trustworthiness on ESConv supporter responses generated by the frozen base (Liu et al., 2021) (random sample of 500 supporter turns), with the same rubric and judges as Table 1; per-dimension scores in the tables below. The Llama-3.1-8B row is the with-prefix variant of Table 1, evaluated here on text-only input.

System	GPT-4o	Claude-4.0-Sonnet	Grok-3
Llama-3.1-8B	5.17	5.91	6.01
GPT-4o (generator)	7.84	6.94	7.78
SCHRONO (w/o prefix)	7.87	6.96	7.80
SCHRONO	**8.10**	**7.30**	**8.09**

Table 14: ESConv generation, GPT-4o judge scores (1 to 10).

System	Tru.	Saf.	Fai.	Pri.	Emp.	Rel.	Eth.	Avg.
Llama-3.1-8B	6.07	4.01	5.16	8.11	4.07	4.56	4.21	5.17
GPT-4o (generator)	7.58	8.40	6.79	8.64	7.51	7.48	8.51	7.84
SCHRONO (w/o prefix)	7.62	8.43	6.82	8.67	7.53	7.51	**8.54**	7.87
SCHRONO	**7.92**	**9.03**	**6.94**	**8.85**	**8.01**	**8.14**	7.81	**8.10**

Table 15: ESConv generation, Claude-4.0-Sonnet judge scores (1 to 10).

System	Tru.	Saf.	Fai.	Pri.	Emp.	Rel.	Eth.	Avg.
Llama-3.1-8B	4.42	6.59	5.46	7.61	5.16	6.55	5.58	5.91
GPT-4o (generator)	6.58	6.97	6.90	8.16	6.99	6.30	6.67	6.94
SCHRONO (w/o prefix)	6.61	6.98	6.92	8.18	7.01	6.32	6.69	6.96
SCHRONO	**6.91**	**7.35**	**7.26**	**8.50**	**7.36**	**6.67**	**7.04**	**7.30**

row is a GPT-4o generator baseline, whose scores are inflated by judge self-preference. Bold: best in column among the listed rows.

What this comparison isolates. ESConv is text-only, so the encoders and the multimodal prefix are inactive and only the base generator and the learned prefix tokens operate. The comparison therefore tests two things at once: whether the frozen backbone still produces coherent supportive responses on a corpus it never saw, and whether the learned prefix helps or hurts when it is the only conditioning signal available. Llama-3.1-8B is included as the autoregressive reference and is evaluated under the same prompt and cleaning protocol, so the gap between it and SCHRONO reflects the base generator rather than a prompt difference.

Results. SCHRONO leads both baselines on every judge, at 8.10 against 7.84 for the GPT-4o generator and 5.17 for Llama-3.1-8B under the GPT-4o judge, with the same ordering under Claude (7.30 / 6.94 / 5.91) and Grok (8.09 / 7.78 / 6.01). The margin over Llama-3.1-8B is large and consistent, which is the part of this table we read as evidence that the frozen masked-diffusion base generalizes off-corpus; the margin over the GPT-4o generator is small (+0.26, +0.36, +0.31) and we treat it as suggestive only, both because the GPT-4o row is scored by a GPT-4o judge that favours its own generations and because a single-cell loss on ethics under that judge (8.51 against 7.81) shows the comparison is not one-sided. The learned prefix improves the no-prefix variant under all three judges (7.87 to 8.10, 6.96 to 7.30, 7.80 to 8.09), so the prefix is not merely an artefact of the MESC training distribution but transfers to a corpus the model was never tuned on.

Table 16: ESConv generation, Grok-3 judge scores (1 to 10).

System	Tru.	Saf.	Fai.	Pri.	Emp.	Rel.	Eth.	Avg.
Llama-3.1-8B	5.66	5.66	6.71	7.15	5.55	6.01	5.33	6.01
GPT-4o (generator)	7.27	8.36	7.10	8.55	7.69	7.27	8.20	7.78
SCHRONO (w/o prefix)	7.28	8.38	7.12	8.57	7.71	7.29	8.22	7.80
SCHRONO	**7.64**	**8.66**	**7.29**	**8.85**	**8.01**	**7.65**	**8.56**	**8.09**

J CONDITIONING DIAGNOSTICS

The four diagnostics are computed on the MESC test set under three conditions: the bidirectional model uses its full Stage-2 conditioning adaptation (prefix tokens and frame-feature projections); the causal model uses the same adaptation re-trained in Llama's embedding space under the identical Stage-2 recipe; and the intervention column applies a causal attention mask to the bidirectional model's own weights and conditioning, with no re-training. Every value is read out from the frozen model's own logits or hidden states, never from a separately trained classifier. The prefix layout is

Table 17: Conditioning diagnostics under three conditions. Columns 1 and 3 are the cross-model comparison, each adaptation trained in its own embedding space; column 2 is the topology intervention, which applies a causal attention mask to the same LLaDA-8B weights and the same learned conditioning without re-training. Higher is better for B4 and lower is better for B3 drift; B2 has no monotone direction, because values approaching 1 indicate the collapse regime rather than better temporal consistency (Section 4.5), so B2 is not bolded. Bold marks the best value in the row under the stated direction. Row meanings follow Section 4.5.

Diagnostic (all in %)	LLaDA-8B bidirectional	LLaDA-8B causal mask	Llama-3.1-8B causal
B1 video removed	21.0	34.7	54.8
B1 audio removed	14.2	22.3	31.7
B1 text removed	90.8	77.3	17.9
B2 inter-frame cosine similarity	0.8980	0.9714	0.6054
B3 KL drift, video dropped ↓	**0.0488**	2.0230	0.9367
B3 KL drift, audio dropped ↓	0.0601	0.4803	0.0333
B3 KL drift, both dropped ↓	**0.0858**	2.1087	0.8477
B4 prefix symmetric ratio ↑	**0.315**	0.137	0.237

$[\, \boldsymbol{q}_1^{\mathrm{v}}, \boldsymbol{q}_2^{\mathrm{v}}, \boldsymbol{f}_1^{\mathrm{v}}, \ldots, \boldsymbol{f}_{40}^{\mathrm{v}}, \boldsymbol{q}_1^{\mathrm{a}}, \boldsymbol{q}_2^{\mathrm{a}}, \boldsymbol{f}_1^{\mathrm{a}}, \ldots, \boldsymbol{f}_5^{\mathrm{a}} \,]$, a total of 49 tokens. Let p_{full} denote the softmax of the next-token logits at the last position under the full prefix. **B1 (cross-modal grounding)**: the fraction of test cases in which zeroing one part of the prefix changes the argmax next-token prediction at the reply position. Zeroing the video sets the two video prefix tokens and the 40 frame tokens, positions $1, \ldots, 42$, to zero, zeroing the audio sets the two audio prefix tokens and the 5 audio tokens, positions $43, \ldots, 49$, to zero, and removing the text replaces the dialogue history with an empty prompt. **B2 (temporal consistency)**: the mean cosine similarity between the L2-normalized last-layer hidden states of all pairs of real (non-zero-padded) video frame tokens, averaged over off-diagonal pairs and over samples. **B3 (modality robustness)**: the KL divergence $\mathrm{KL}(p_{\mathrm{full}} \,\|\, p_{\mathrm{drop}})$ between the next-token distributions with the full prefix and with the video, audio, or both zeroed. Lower indicates higher robustness. **B4 (position-agnostic influence)**: for every non-padding text position j, the KL divergence between the output distribution at position j with the full prefix and with the prefix zeroed. Per-position influences are averaged over samples, the mean profile is split into its first and second halves, and the reported ratio is $\min(\bar{a}, \bar{b}) / \max(\bar{a}, \bar{b})$ of the two half means. Higher means the prefix conditions early and late positions more uniformly.

Same-checkpoint topology intervention. The intervention changes exactly one thing: the attention mask of the frozen backbone. The checkpoint, the tokenizer, the prefix layout, the frame and audio token counts, the learned conditioning inputs, the test examples, and the read-out of logits and hidden states are identical to the bidirectional condition. No parameter is updated and no re-training is performed, so the intervention recomputes B1–B4 from the same weights under the intervened information flow; in the released code it is a single evaluation-time switch, with the data split and seed held fixed.

Reading each diagnostic. B1 is a discrete intervention count: for each test case we record whether zeroing video, audio, or the text of the dialogue history changes the argmax next-token prediction at the reply position, and report the fraction of cases where it does. A high value means the model's prediction depends on that carrier, so the direction of interest differs per row: for video, higher for the causal model means the bidirectional model is less disturbed by losing a modality, whereas for text the high bidirectional value (90.8%) shows text dominance rather than robustness, and the text row is therefore not evidence in favour of either topology.

B2 is a representation-level statistic with no monotone quality direction: a pairwise cosine near 1 is reached both by mutually consistent frames and by frames collapsed onto the same vector, and Proposition B.3 shows the latter is what uniform attention produces. This is why we pair B2 with the linear probe (Section 4.5) rather than reading it alone, and why Section 4.5 finds no rank agreement between B2 and judge quality across systems while B3 and B4 both do.

B3 is a distributional distance between the model's own next-token distributions with and without a modality, so it is read out from logits and requires no trained classifier. The video row separates the conditions most sharply (0.0488 bidirectional, 2.0230 intervened, 0.9367 causal Llama), the both-dropped row follows the same order, and the audio row does not, which is consistent with the modality ablation in Appendix M: when a modality carries little information, dropping it barely moves any of the three conditions, so the audio ordering carries no signal. B4 measures how evenly

the prefix conditions the text by splitting the per-position influence profile into halves and reporting $\min / \max$ of the two half means; a value below 1 means the profile is front-loaded, and the ordering here (0.315 bidirectional, 0.137 intervened, 0.237 causal Llama) places the intervened model below the causal backbone, i.e. the causal mask makes prefix influence more position-dependent than causal pretraining alone.

Intervention scope. The intervention changes exactly one thing: the attention mask. The checkpoint, tokenizer, prefix layout, frame and audio token counts, learned conditioning inputs, test examples, and the read-out of logits and hidden states are identical to the bidirectional condition, no parameter is updated, and the data split and seed are held fixed, so every difference in Table 17 is attributable to the mask. The cost of this control is that the intervened model is out of distribution with respect to its pretraining objective, so its diagnostics bound how far the learned conditioning relies on global visibility rather than estimating the performance of a causally pretrained system; the cross-backbone column with Llama-3.1-8B provides that second, complementary reading, at the cost of confounding topology with backbone family.

K Diagnostics across independently built systems

Table 3 in the main text gives the per-system values behind Section 4.5: B2, B3, and B4 computed for each of the six systems that have LLM-judge means in Table 1, using each system's own generator and its own learned conditioning, together with the Spearman rank correlation between each diagnostic and the judge mean. We give the reasoning behind the ordering, why B2 is a null, and the uncertainty on the correlations here.

Why the ordering is not an artifact. Two of the six systems are ablations of our own model, so a rank agreement computed only over those would be circular. Three of the remaining four are other authors' systems (Qwen2-7B, both Llama-3.1-8B variants, and MultiMood SFT+RL), and B3 places all of them in exactly the order their judge scores do, while B4 matches on all but the two Llama variants, which differ by 0.03 on the judge mean and 0.014 on the ratio. The correlation is therefore not produced by comparing our own architecture against itself.

Why B2 is reported as a null. Raw inter-frame similarity has no monotone quality direction, and the table shows why: the highest B2 value in the table (MultiMood SFT+RL at 0.9703) belongs to a mid-ranked system, while the bidirectional SCHRONO model sits at 0.8980 and the intervened model at 0.9714 (Table 17) despite the two scoring very differently on the judges. This is consistent with the reading in Section 4.5: similarity near 1 means consistency only when the representations remain decodable, which the linear probe tests directly.

Uncertainty on the correlations. With $n = 6$ the exact permutation p-values in Table 3 are the appropriate test. Resampling the six systems with replacement 20,000 times and recomputing the rank correlation against the judge mean gives $\rho = -1.00$ with a 95% interval of $[-0.94, -0.43]$ for B3 and $\rho = +0.94$ with a 95% interval of $[+0.54, +1.00]$ for B4. Both intervals approach parity at the lower end: the direction of the association is stable under resampling, but the intervals do not exclude a moderate correlation, and B3's point estimate sits at the boundary of the attainable range. The rank agreement supports the diagnostics; it does not calibrate them.

L Human evaluation protocol

The human study follows the pairwise-preference design of BiGraph-Diffuse (Cheng et al., 2026). One hundred test dialogues are used; for each dialogue, raters see the anonymized client context and two de-identified counselor responses (SCHRONO versus a baseline, order randomized), and indicate a win, tie, or loss. Raters are blind to the system identity, and only short text excerpts are shown (no client-identifying audio or video).

Lay panel. Fifty lay participants with conversational experience rate response pairs along four dimensions: comprehensiveness (depth and accuracy of understanding of the client's concerns), professionalism (structuring and technique), authenticity (naturalness and genuineness), and empathy (emotional attunement). Inter-rater agreement among the lay raters was substantial (Kendall's $W = 0.71, p < 0.001$).

Table 18: Lay and expert preference evaluation: win rate of SCHRONO (%) in pairwise win/tie/lose comparisons on 100 test dialogues. Lay panel (top, 50 participants): Com. comprehensiveness, Pro. professionalism, Att. authenticity, Emp. empathy. Expert panel (bottom, 5 licensed clinical psychologists): Emp&Ins. empathy and insight, Sup&Aut. support and autonomy, Att&Pre. attunement and presence, Saf&Bou. safety and boundaries. The Llama-3.1-8B rows are the with-prefix variant of Table 1. Overall rates and Wilson intervals are discussed in Section 4.6.

Lay panel: pairwise comparison	Com.	Pro.	Att.	Emp.	Overall
SCHRONO vs. MultiMood (SFT+RL)	66.8	65.9	70.4	69.3	68.1
SCHRONO vs. Llama-3.1-8B	75.7	74.2	78.6	77.9	76.6
Expert panel: pairwise comparison	**Emp&Ins.**	**Sup&Aut.**	**Att&Pre.**	**Saf&Bou.**	**Overall**
SCHRONO vs. MultiMood (SFT+RL)	71.4	69.8	70.6	74.2	71.5
SCHRONO vs. Llama-3.1-8B	78.9	77.3	79.5	82.1	79.4

Table 19: Agreement between LLM-judge scores and human preferences. The judge margin is the difference between SCHRONO's mean trustworthiness score and the baseline's; the human columns report SCHRONO's pairwise win rate.

System pair	Judge margin	Lay win	Expert win
SCHRONO vs. MultiMood SFT+RL	8.03 vs. 7.15 (+0.88)	68.1%	71.5%
SCHRONO vs. Llama-3.1-8B (w/ prefix)	8.03 vs. 6.03 (+2.00)	76.6%	79.4%

Expert clinical assessment. Five licensed clinical psychologists, representing CBT, humanistic, and psychodynamic orientations, rate the same pairs along four clinical dimensions: empathy and insight, support and autonomy, attunement and presence, and safety and boundaries. Inter-rater reliability among the experts was high (ICC(2,1) $= 0.81$). Confidence intervals are Wilson intervals over the 100 test dialogues, computed from the reported win rate for each comparison, which is the conservative choice for proportions near 0.7 at this sample size; the intervals are [58.4, 76.4] and [67.4, 83.8] for the lay comparisons and [62.0, 79.4] and [70.5, 86.2] for the expert comparisons, so every interval lies above 50%. Those intervals treat each dialogue as one independent observation and so do not capture the dependence induced by the same rater scoring many dialogues; a rater-level analysis would widen them, and we report the dialogue-level intervals as the narrower bound.

The study was conducted as an anonymous questionnaire with informed consent, voluntary participation, and the right to withdraw at any time; it collected no personal, medical, or biometric data beyond the preference ratings, and the ratings were stored and analyzed in aggregated, anonymized form (see the Ethics statement at the end of the paper).

Agreement with LLM judges. The human preference data aligns in direction with the LLM-judge scores (Table 19). SCHRONO's judge margin over MultiMood SFT+RL (8.03 vs. 7.15, +0.88) corresponds to a 68.1% lay and 71.5% expert win rate, and its larger margin over Llama-3.1-8B (8.03 vs. 6.03, +2.00) corresponds to higher win rates (76.6% lay, 79.4% expert). With two baseline comparisons the agreement is qualitative rather than a correlation coefficient, but the direction is consistent across both rater groups and both baselines.

M MODALITY CONTRIBUTION

Table 20 reports the modality ablation behind Appendix O, evaluated on the MESC test set with all three encoders trained under the standard two-stage schedule and one input channel removed at inference.

Reading the table. Text is the strongest single carrier (56.14 alone, against 46.24 for video and 41.25 for audio), which matches the B1 text-removal row in Table 17. Each modality also carries complementary information, since pairing any two beats either alone by a wide margin: video plus text reaches 64.70 and audio plus text 61.23, both far above the 56.14 of text alone, and video plus audio reaches 57.26 with no text at all. The full combination is best at 68.70, so the three channels are not redundant. The ordering of the single-modality rows (text ¿ video ¿ audio) is also what the diagnostic suite independently implies: video removal perturbs the reply prediction far more than audio removal under every condition in Table 17, and the audio channel is the one whose diagnostic ordering is uninformative. The asymmetry between video and audio motivates the limitation we

Table 20: Modality contribution on MESC user emotion recognition (accuracy), mean over three runs on the full test set; full multimodal input is best.

Input combination	Emotion accuracy
full [video + audio + text]	**68.70**
video + text	64.70
audio + text	61.23
video + audio	57.26
text only	56.14
video only	46.24
audio only	41.25

Table 21: Component ablations on MESC (accuracy, final-epoch weights on the full test set), ablating one design choice at a time from the full model.

Design choice	Variant	Emotion	Strategy
Attention pooling	ours	**68.70**	**50.52**
	mean pooling	67.4	49.3
Late fusion	ours	**68.70**	**50.52**
	single head	67.9	49.6
Classifier projectors	ours	**68.70**	**50.52**
	shared	67.8	49.4
Curriculum	ours	**68.70**	**50.52**
	single-stage	44.6	33.8

state in the Discussion: improving audio fusion is the most direct route to further gains, because audio is the weakest link in an otherwise well-fused stack.

N COMPONENT ABLATIONS

Table 21 ablates one design choice at a time from the full model (accuracy, final-epoch weights on the full test set), as analyzed in Appendix O. Ablation variants are single runs; the run-to-run std of the full model is $\pm$ 0.9 on emotion accuracy and $\pm$ 1.1 on strategy accuracy (Table 2), so deltas inside that band are indicative rather than established.

Variants. *Mean pooling* replaces the learnable attention pooling of Section 3.2 with a uniform average over frame and audio features, keeping every other component identical, which tests whether attention over frames is doing work beyond aggregation. *Single head* removes the late fusion of equation 4 and reads the prediction from the multimodal head alone, testing whether the text-only head adds complementary signal. *Shared projectors* drops the frozen Stage-1 projector clones so that the classification and generation pathways share one projector that both objectives update, which is the configuration Proposition B.1 predicts should let the generation gradient dominate. *Single-stage* trains the classifier and the generator jointly from scratch with no Stage-1 adaptation and no curriculum, which isolates the contribution of the two-stage schedule.

What the ordering shows. The curriculum is by far the largest factor: removing it costs 24.1 points of emotion accuracy and 16.7 points of strategy accuracy, an order of magnitude more than any architectural variant, and the gap is far outside the run-to-run std. Among the architectural choices the ordering is consistent but small, at 1.3, 0.9, and 0.8 points respectively, all within or barely above the noise band, so we report them as directional rather than as separately established effects. The shared-projector result is the one we lean on most, because it is the empirical counterpart of the exact decoupling argument: it is the only architectural variant whose removal is predicted a priori to hurt, and it does.

Feature ablation. Separately from the design-choice ablations, zeroing the video and audio features while keeping the learned prefix tokens in place drops the MESC judge mean from 8.03 to 7.8, which shows that the sensory content rather than the learned tokens alone is what shapes generation. This is a different intervention from the modality ablation of Table 20: there we remove an input to a trained classifier, here we remove the conditioning signal while leaving the prefix structure intact, so the two together separate the contribution of the learned tokens from that of the features they carry.

Table 22: Generator ablation on the MESC test responses. All rows use the same frozen LLaDA-8B weights. "Adapter" is the merged BiGraph-Diffuse LoRA adapter; "prefix" the learned Chrono conditioning tokens; "sensory" the per-frame video and audio tokens; "mask" the attention mask of the frozen backbone. Judge mean is the average over the three judges of Table 1 on the seven trustworthiness dimensions.

Configuration	Adapter	Prefix	Sensory	Judge mean
LLaDA-8B (base)	–	–	–	5.60
+ adapter	✓	–	–	7.55
+ prefix	✓	✓	–	7.68
+ sensory (SCHRONO)	✓	✓	✓	**8.03**
SCHRONO, causal mask	✓	✓	✓	7.10

O MODALITY AND COMPONENT ANALYSIS

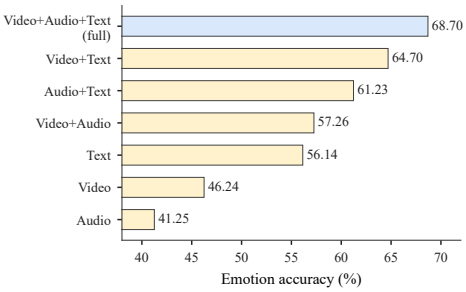

Figure 4: Modality contribution on MESC user emotion recognition (accuracy). Exact values in Appendix M.

Ablating the input modalities (Figure 4), the full combination is best (68.70): video plus text (64.70) and audio plus text (61.23) each improve over text alone (56.14), so both extra channels carry complementary information while text remains the primary carrier. The component ablations (Appendix N) remove one design choice at a time: attention pooling outperforms mean pooling by 1.3 points on emotion, removing the late fusion of the text-only head (Eq. 4) costs 0.8, and sharing the classifier projectors rather than the frozen clones costs 0.9. Two of these deltas lie inside the run-to-run std of the full model ($\pm$ 0.9 emotion, $\pm$ 1.1 strategy, Table 2) and the pooling gain only slightly exceeds it, so we read all three as consistent trends from single-run ablations. The two-stage curriculum is the largest contributor: joint training from scratch drops emotion accuracy to 44.6 and strategy accuracy to 33.8, because the diffusion loss over the full answer dwarfs the small auxiliary losses. Zeroing the video and audio features while keeping the learned prefix tokens also drops the MESC judge mean from 8.03 to 7.8, so the sensory content shapes generation rather than the learned tokens alone.

P GENERATOR ABLATION

Table 22 decomposes the generation result of Section 4.2 into the contributions of the frozen adapter, the learned prefix, and the sensory conditioning. Because the backbone is never updated, the rows correspond to inference configurations of the same weights, so the differences are attributable to what is fed to the generator rather than to what was trained. The final row reports the same fully conditioned configuration with the bidirectional attention mask replaced by a causal one, which isolates the mask from the conditioning stack.

How to read this. The largest single step is the adapter itself (+1.95), the component that adapts a general diffusion language model to counseling text and that is shared with BiGraph-Diffuse (Cheng et al., 2026). The two components SCHRONO adds are smaller: the prefix contributes +0.13, consistent with the 7.93 to 8.03 comparison in Section 4.2, and the sensory conditioning contributes +0.35, the part of the result that depends on the client's video and audio rather than on text alone. The causal-mask row is the one that speaks to information flow rather than to conditioning content: with the same weights, the same prefix, and the same per-frame tokens, substituting the mask lowers the judge mean from 8.03 to 7.10, a drop of 0.93 that is larger than the 0.35 contributed by the sensory conditioning and larger than the 0.13 contributed by the prefix. Since the intervened model is out of distribution relative to its bidirectional pretraining, part of that drop reflects the mismatch rather than the topology, and the row bounds rather than estimates the effect of causality on generation. It does show that within a single set of weights the mask is not a neutral read-out choice: changing it alone moves generated quality by more than the entire learned conditioning stack does.

Table 23: Cross-corpus emotion recognition: accuracy (%) without target-corpus adaptation and after fine-tuning, on the shared classes of IEMOCAP and MELD. IEMOCAP reports the mean ± std over five leave-one-session-out folds; MELD is a single evaluation on the six classes shared with MESC.

	IEMOCAP	MELD (6 classes)
No target-corpus adaptation	44.8 ± 3.6	51.05
Fine-tuned on target	57.9 ± 2.7	–

Q CROSS-CORPUS TRANSFER DETAILS

Results. The decoupling claim has a testable consequence beyond MESC: because the backbone is never retrained and the sensory machinery is a small additive stack, neither should be bound to the corpus it was trained on. We evaluate the MESC-trained sensory machinery and the base generator on IEMOCAP (without target-corpus adaptation, and then fine-tuned) and on MELD (without target-corpus adaptation).

Emotion recognition on IEMOCAP. IEMOCAP (Busso et al., 2008) is a dyadic multimodal emotion corpus whose labels come from a different inventory than MESC, testing both cross-corpus and cross-setting transfer. Applying the MESC-trained machinery without any adaptation gives 44.8 ± 3.6% on the four shared classes, 19.8 points above chance but well below in-corpus accuracy; fine-tuning with the emotion head re-initialized reaches 57.9 ± 2.7%, 13.1 points higher, with non-overlapping intervals. The unadapted figure shows partial transfer under a large cross-corpus shift, and the fine-tuned figure shows the recipe retains learnability on a second corpus.

Emotion recognition on MELD without target-corpus adaptation. On the six classes shared with MESC (2,329 of the 2,610 test utterances, surprise excluded) the same machinery reaches 51.05% accuracy and 40.22% weighted F1. The audio projector received MELD pretraining in Stage 0, so the genuinely unseen components here are the video pathway and the fusion level.

IEMOCAP protocol. IEMOCAP (Busso et al., 2008) provides utterance-level categorical emotion labels; we use the four classes shared with MESC (angry, happy, sad, and neutral), taken directly from the consensus majority label in EmoEvaluation, giving 4,490 utterances (angry 1,103, happy 595, sad 1,084, neutral 1,708). Excited is excluded rather than merged with happy, because the two labels differ in arousal and the raw four-class consensus protocol is the most common in the literature. Evaluation follows leave-one-session-out cross-validation (5 folds): the ten speakers across the five sessions are disjoint, so the speakers of a held-out session were never seen during training, which gives speaker independence with five folds rather than ten. We report the mean accuracy and standard deviation over the five folds.

MELD protocol. MELD (Poria et al., 2019) annotates seven emotions; MESC and MELD share six of them, and we exclude surprise (present in MELD but not in MESC), evaluating on 2,329 of the 2,610 test utterances. Because the audio projector was pretrained on MELD in Stage 0, these numbers should be read with that caveat: the video pathway (AffectNet-pretrained) and the fusion and classification level are the components that were never trained on MELD.

Table 23 summarizes the accuracy without target-corpus adaptation and after fine-tuning for both evaluations.

R EFFICIENCY ANALYSIS

The frozen-backbone design keeps the training and serving cost close to that of the underlying diffusion language model. The LLaDA-8B backbone with its merged counseling adapter contributes roughly 8B parameters that are never updated. In Stage 1, the SigLIP (878M) and WavLM (94M) encoders are unfrozen together with the sensory heads, giving about 0.97B trainable parameters while the encoders adapt to the multimodal semantics. In Stage 2, the encoders are frozen again and only the projectors, attention pooling, prefixes, and heads are trained, roughly 20M parameters, about 0.25% of the base model. The full pipeline runs on a single NVIDIA RTX 6000D (84 GB) in about 24 GPU-hours. At inference, generation runs the masked-diffusion sampler for 256 denoising steps, with the steps executed block-parallel across positions; an autoregressive model decodes the same number of output tokens with one forward pass per token, so both regimes execute roughly the same number of sequential steps. In wall-clock terms the two are close in practice: on a single RTX

Table 24: Efficiency summary.

Component	Value
Backbone (frozen throughout)	LLaDA-8B + BiGraph adapter (8B)
Stage 1 trainable	$\approx$ 0.97B (SigLIP 878M + WavLM 94M + heads)
Stage 2 trainable	$\approx$ 20M ($\approx 0.25\%$ of base)
Training hardware	$1\times$ NVIDIA RTX 6000D (84 GB)
Training time	$\approx$ 24 GPU-hours
Generation steps	256 (block-parallel)

6000D, generating the complete MESC test responses (454) takes approximately 90 minutes with SCHRONO and approximately 80 minutes with Llama-3.1-8B under identical batching, a gap of about 12%. The theoretical per-step cost of full-sequence attention is offset by the parallel decoding of the diffusion sampler, whereas the autoregressive baseline decodes serially with a growing key-value cache. Classification is a single forward pass through the encoders and heads. Table 24 summarizes these figures.

Where the cost sits. Two properties of the design drive these figures. First, freezing the backbone means Stage 2 trains roughly 0.25% of the base parameters, so training cost is dominated by the Stage 1 encoder adaptation rather than by the language model. Second, the diffusion sampler trades sequential depth for per-step cost: it executes about as many sequential steps as autoregressive decoding of the same output length, but each step scores every position in parallel, which is why the measured wall-clock gap is 12% rather than the ratio of parameter counts. Serving cost therefore tracks the base model, and the sensory stack adds a fixed encoder pass per turn.

Headroom. The two largest reducible costs are the diffusion step count and the Stage 1 encoder adaptation. Distillation or early-exit over the denoising trajectory would cut the 256 steps directly, and the encoders could be frozen earlier if a smaller Stage 1 budget proved sufficient. Neither change affects the decoupling property the method relies on, since both leave the backbone untouched.