PROSODIC VIDEO NARRATION FROM STREAMING AUDIO-VISUAL EVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Compelling spoken narration requires not only the right words, but also prosody that reflects the significance of unfolding events. Existing streaming video models focus on *what* to say, while text-conditioned speech models lack the perceptual context needed to determine *how* to say it. We introduce *Prosodic Video Narration*, a task of continuously generating spoken descriptions whose content and prosody are grounded in streaming audio-visual events. To support this task, we construct a sports commentary dataset with aligned video, environmental audio, commentary text, and word-level prosodic annotations. We further propose PROSODYMLM, a streaming multimodal language model that models discrete word-level prosody from audio-visual context. Our design enables efficient autoregressive prosody prediction through lightweight adapters and prosody injection, while a two-stage training strategy transfers expressive speech capabilities before grounding them in multimodal events. Across quantitative metrics and a user study, PROSODYMLM generates narration that better aligns with reference speech and underlying events than existing baselines, improving pitch, speaking rate, and emotional alignment while maintaining competitive energy alignment, with particularly strong gains during highly expressive moments.
Anonymous demo page: `https://prosodymlm.github.io/`

1 INTRODUCTION

The 2026 World Cup offered countless reminders that compelling narration is about more than saying the right words. When a striker breaks through the defense in stoppage time and scores a decisive late game winner, a commentator may stretch the scorer's name into an ecstatic shout, raising their pitch and energy to amplify the significance of the moment. Such context-aware prosody matters far beyond sports. For a visually impaired user wearing smart glasses, an assistive personal agent should urgently shout "STOP!" when a vehicle is rapidly approaching, but calmly mention a cat crossing the street; and an interactive household robot may similarly speak with urgency when it sees smoke and hears a fire alarm, but remain calm when describing routine events.

However, existing streaming video captioning and narration systems (Xu et al., 2026; Chen et al., 2025) largely focus on generating the right words, while expressive speech models (Qian et al., 2025; Hu et al., 2026) model how to say them but are only conditioned on text or speech context. As a result, their narration can sound flat or mismatched to the scene. As illustrated in Fig. 1, when Messi simply walks into position amid ordinary stadium ambience, the narration should remain calm. Moments later, when he takes a shot and scores as the crowd erupts in cheers, shouts, and applause, the delivery should become dramatically more expressive. Effective narration therefore requires continuously interpreting streaming visual and acoustic events and dynamically adapting its prosody to the significance of each moment.



Figure 1: Illustration of streaming *Prosodic Video Narration* leveraging audio-visual cues.

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.

Towards this goal, we introduce *Prosodic Video Narration*, a new task of continuously generating spoken descriptions with aligned prosody from streaming visual and audio inputs. To address this task, we propose PROSODYMLM, a streaming multimodal language model that predicts word-level prosody for incoming commentary while grounding its delivery in current and historical audio-visual context. PROSODYMLM introduces lightweight prosody modeling on top of a pretrained streaming backbone, injecting quantized prosodic representations directly into corresponding word representations to preserve word-level alignment and maintain coherence with previously generated prosody, without increasing sequence length. We train the model in two stages, first distilling expressive prosodic patterns from large-scale pseudo-labeled text data and then grounding them in streaming visual and acoustic events through multimodal fine-tuning. This design enables efficient adaptation while preserving the pretrained model's streaming language capabilities.

We also curate a new dataset for this task from existing sports videos, comprising 420 hours of aligned visual content, environmental audio, spoken commentary, and word-level prosodic annotations. Evaluated on this dataset, PROSODYMLM generally outperforms strong baselines across objective prosody metrics, generating commentary with more appropriate expressive intensity and stronger alignment between prosody and the underlying audio-visual events. A user study further confirms these improvements, with participants generally favoring our generated commentary in terms of both overall quality and event-prosody alignment.

In summary, our main contributions are as follows:

- We introduce *Prosodic Video Narration*, a new task of continuously producing spoken narration with word-level prosody grounded in streaming audio-visual events.
- We curate a sports video dataset with aligned visual events, environmental audio, spoken narration, and word-level prosodic attributes, and propose PROSODYMLM, a streaming multimodal language model for context-aware prosody generation.
- We show that PROSODYMLM outperforms strong baselines on both objective metrics and human evaluation, generating coherent and natural narration whose pitch, speaking rate, and emphasis better align with streaming visual events and environmental sounds, while maintaining competitive energy alignment.

2 RELATED WORK

Multimodal Language Models. Recent advances in multimodal language models have shown that combining visual encoders with pretrained language models enables strong image and video understanding. LLaVA (Liu et al., 2023) and the Qwen-VL family (Bai et al., 2023; Wang et al., 2024; Bai et al., 2025b;a) use visual instruction tuning to align visual features with the language-model embedding space, while Video-LLaMA 2 (Cheng et al., 2024) extends this paradigm to temporal video understanding. OneLLM (Han et al., 2024) supports diverse multimodal inputs through a universal projection module with dynamic routing, and more recent methods such as Qwen3-Omni (Xu et al., 2025) and Gemma 4 (Team et al., 2026) further unify modeling across text, vision, and audio. Similar to these approaches, we condition a pretrained language model on visual and acoustic representations; however, our goal is not only to generate the right text, but also to predict fine-grained prosody that dynamically aligns with the underlying audio-visual events.

Streaming Video Language Models. To support real-time, long-form video understanding, VideoLLM-online (Chen et al., 2024) introduces a streaming objective that processes incoming frames and generates temporally aligned responses online. LiveCC (Chen et al., 2025) trains video-language models on large-scale videos paired with ASR transcripts, interleaving frames and words by timestamp for fine-grained temporal grounding and applications such as play-by-play sports commentary. StreamingVLM (Xu et al., 2026) further extends the effective video context through an efficient KV-cache mechanism, preserving history while processing long video streams with bounded computational overhead. Building on these advances, we adopt StreamingVLM as the backbone of PROSODYMLM and extend it from streaming text generation to prosody-aware generation grounded in both visual events and environmental audio.

Speech Generation Models. Generating realistic and expressive speech with controllable prosodic variation has become an active research area. Discrete speech tokenization (Guo et al., 2025; Ahasan et al., 2024) converts continuous waveforms into discrete representations, enabling speech language

models to understand and generate speech using sequence modeling. Building on these representations, encoder–decoder (Li et al., 2023) and language-model-based approaches (Borsos et al., 2023; Zhang et al., 2023) model various attributes of expressive speech. ProsodyLM (Qian et al., 2025) provides fine-grained control through word-level tokens representing duration, pitch, and energy, while Qwen3-TTS (Hu et al., 2026) supports speech control through natural-language instructions and reference audio. Recent work has also incorporated visual context: MoshiVis (Royer et al., 2026) equips a speech language model with visual understanding, while VoiceCraft-Dub (Sung-Bin et al., 2025) uses facial cues for expressive speech and lip synchronization. In contrast to these speaker-centric settings, we ground prosody in scene-level visual events and environmental sounds, allowing expressive delivery to adapt to what is unfolding in the video.

AI for Sports Applications. Sports video has long served as a challenging testbed for video understanding, with prior work studying event detection (Giancola et al., 2018; Xu et al., 2024), player tracking (Cui et al., 2023; Somers et al., 2024), multimodal reasoning (Rao et al., 2025; Xia et al., 2026), and commentary generation (Rao et al., 2024; Chen et al., 2025; Xu et al., 2026). While prior commentary systems primarily focus on generating what to say from video, we introduce a new task of *Prosodic Video Narration*, which additionally models how narration should be spoken based on unfolding audio-visual events, and use sports videos as our primary testbed and application.

3 The Prosodic Video Narration Task

3.1 Task Definition

We introduce *Prosodic Video Narration*, a new task of generating coherent, continuous spoken commentary from streaming audio-visual observations, with prosody appropriate to the visual content and surrounding acoustic atmosphere. For a streaming video input, we divide the stream into segments and define the recent audio-visual context as:

$$\mathcal{H}_{t-k:t} = \{(\mathbf{V}_{t-k}, \mathbf{A}_{t-k}), \ldots, (\mathbf{V}_{t-1}, \mathbf{A}_{t-1}), (\mathbf{V}_t, \mathbf{A}_t)\}, \tag{1}$$

where $\mathbf{V}_t$ and $\mathbf{A}_t$ denote the visual frames and environmental audio observed during the t-th segment. At time step t, the model has access to the current segment, a bounded history of the previous k multimodal observations, and its previously generated commentary.

Given these inputs, the model should produce spoken commentary for the current segment:

$$\widehat{\mathbf{S}}_t = f_\theta\left(\mathbf{V}_t, \mathbf{A}_t, \mathcal{H}_{\text{t-k:t-1}}, \widehat{\mathbf{S}}_{\text{t-k:t-1}}\right), \tag{2}$$

where $\mathcal{H}_{\text{t-k:t-1}}$ denotes the bounded historical audio-visual context and $\widehat{\mathbf{S}}_{\text{t-k:t-1}}$ denotes the generated commentary that corresponds to the previous segments. We focus on the prosody-generation component of this task: given continuously arriving commentary text, the model predicts its word-level prosody from the streaming multimodal context and previously generated commentary and prosody.

Prosody. Prosody describes how an utterance is delivered beyond its literal linguistic content. It encompasses temporal, pitch, and energy characteristics of speech, including pauses, phoneme and word duration, speaking rate, intonation, pitch variation, and vocal intensity. These characteristics jointly convey the rhythm, emphasis, and emotional expressiveness of spoken commentary. In our task, appropriate prosody should be grounded in both visual events and environmental audio. For example, increasing pitch variation, speaking rate, or vocal intensity during a salient event accompanied by loud cheering, while maintaining a flatter and calmer delivery during less eventful moments.

Discrete representation of prosody. To enable an LLM to take and generate prosody attributes together with words, we adopt the word-level discrete representation (Shechtman & Fernandez, 2023; Qian et al., 2025). The i-th word in segment t is associated with

$$\mathcal{W}_{t,i} = \left(p^{\text{sil}}_{t,i}, w_{t,i}, p^{(1)}_{t,i}, p^{(2)}_{t,i}, p^{(3)}_{t,i}, p^{(4)}_{t,i}, p^{(5)}_{t,i}\right), \tag{3}$$

where $p^{\text{sil}}_{t,i}$ is the silence duration before $w_{t,i}$, and $p^{(1)}$–$p^{(5)}$ represent average phoneme duration, log-F_0 range, median log-F_0, log-F_0 slope, and log-energy, respectively. Each attribute is normalized

with corpus-level statistics, clipped, and quantized into one of Q values. For each prosodic attribute $r \in \mathcal{R} = \{\text{sil}, 1, \ldots, 5\}$,

$$p_{t,i}^{(r)} \in \mathcal{V}_P = \{0, \ldots, Q-1\}. \tag{4}$$

A speech decoder (Li et al., 2023) converts the commentary-prosody sequence into narration.

3.2 Dataset Construction

Existing streaming video datasets pair video and text, but do not provide the fine-grained prosodic supervision needed to study prosody grounded in unfolding audio-visual events. We focus on sports videos, which are well studied and show a clear link between events and prosody. We therefore construct Inf-Stream-Prosody, a multimodal sports-commentary dataset that aligns video, environmental audio, commentary text, speech, and word-level prosody.

We build Inf-stream-prosody from FIFA and NBA broadcasts in Inf-Stream-Train (Xu et al., 2026). Our source data contain 234 videos (420.2 hours). We separate spoken commentary from environmental sound using `sam-audio-base` (Shi et al., 2025), transcribe the commentary with a word-level ASR model (Sekoyan et al., 2025), and align individual words to speech using a StyleTTS2-based aligner (Li et al., 2023). From these alignments, we follow Qian et al. (2025) and extract the preceding-silence duration and five word-level prosodic attributes defined in Sec. 3.1 and quantize them into discrete prosody tokens. We then segment the commentary into sentence-level utterances and pair each utterance and its prosody with the temporally corresponding video and environmental audio. After filtering corrupted segments, the resulting dataset contains approximately 220K aligned utterances. We use 90% of videos for training and reserve 10% videos for testing.

For out-of-domain evaluation, we additionally construct FIFA2026 from FIFA 2026 broadcasts using the same preprocessing pipeline. Although FIFA2026 also consists of soccer commentary, it covers different matches and broadcast sources from those used in the training set, providing a separate evaluation domain. The dataset contains approximately 4K aligned commentary utterances and is used exclusively for evaluation. Additional details on data collection and preprocessing are provided in Appendix A.1.

4 The ProsodyMLM Model

In this section, we introduce ProsodyMLM, illustrated in Fig. 2, for streaming prosody generation that predicts word-level prosodic attributes from incoming commentary while grounding them in the corresponding visual and acoustic context, as well as maintaining coherence with previously generated commentary. Our approach consists of three components: (1) a multimodal encoder that preserves fine-grained representations of the current segment while compressing historical observations to maintain temporal context; (2) a lightweight adapter that extends a pretrained VLM to autoregressively predict word-level prosodic attributes; and (3) a two-stage training strategy that efficiently equips the model with prosody generation and multimodal grounding capabilities.

4.1 Streaming Multimodal Encoder

To provide contextual evidence for prosody prediction, we use a multimodal encoder to map visual and acoustic signals into the language-model space. Retaining all historical visual and audio tokens, however, would rapidly consume the available context window. We therefore adopt an asymmetric representation: historical observations are compressed into a single token per segment via mean pooling, while fine-grained tokens are retained for the current segment. This design preserves compact historical context for temporally coherent prosody generation while maintaining detailed current observations for event-conditioned prosody prediction.

Because the backbone model already extracts good visual representations, we freeze both its vision encoder and visual projector during training. This also ensures that text generation can reuse these features. For environmental audio, we use a frozen audio encoder and introduce a trainable audio projector that maps its outputs to the language model's hidden dimension. The audio projector is optimized during multimodal fine-tuning, while the pretrained language-model backbone remains frozen. This design enables the model to incorporate acoustic information with minimal modification to its backbone.

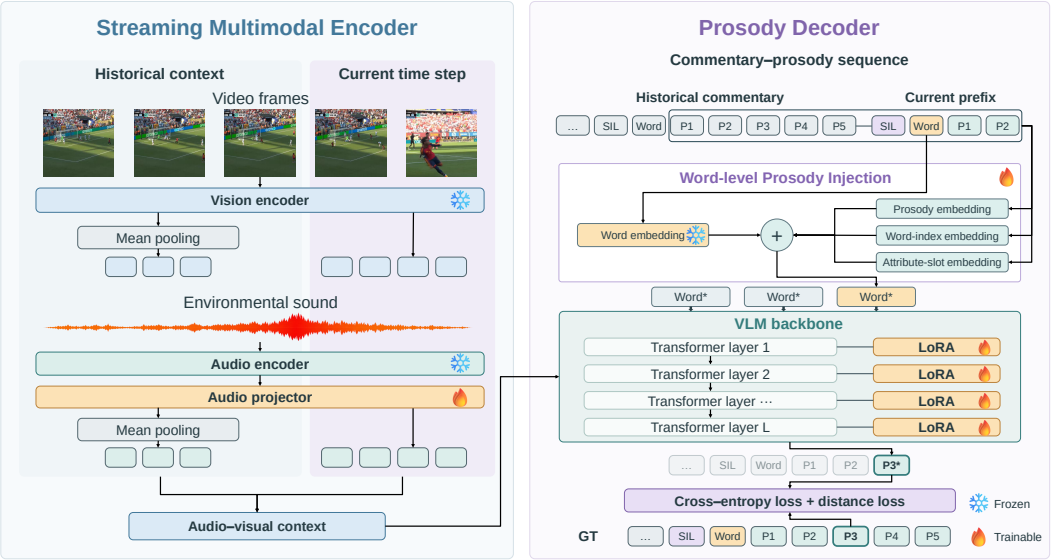

Figure 2: **Overview of ProsodyMLM**. We extend a pretrained VLM with lightweight adapters and word-level prosody injection for streaming prosody generation. Given continuously arriving commentary text aligned with current and historical visual and acoustic context, the model autoregressively predicts word-level prosodic attributes. Previously generated commentary and prosody are retained as context to maintain coherent delivery across segments.

4.2 Prosody Decoder

Given the incoming commentary and streaming multimodal context, the prosody decoder autoregressively predicts the word-level prosodic attributes described in Sec. 3.1. Each prediction is conditioned on both the current word and previously generated prosody, allowing the model to adapt to unfolding events while maintaining a coherent prosodic trajectory over time.

Word-level Prosody Injection. To represent quantized prosody without modifying the pretrained text vocabulary or positional structure, we introduce a numerical vocabulary $\mathcal{V}_P = \{0, \ldots, Q-1\}$ that is disjoint from the text vocabulary $\mathcal{V}_T$. We keep the pretrained text embedding table $\mathbf{E}_T$ frozen and learn a prosody embedding table $\mathbf{E}_P \in \mathbb{R}^{Q \times d}$, where d denotes the hidden dimension of the language-model backbone. As the prosodic attributes are highly correlated, we share the $\mathbf{E}_P$, which also encourages a common numerical representation and keeps the prediction space compact.

A quantized value alone does not specify either the word to which it belongs or the prosodic attribute that it represents, so we encode prosody along two structural dimensions: a word index i and an attribute slot r. For attribute $r \in \mathcal{R}$ of word i, we define

$$\mathbf{z}_{i,r} = \mathbf{E}_P[p_i^{(r)}] + \mathbf{E}_{\text{wpos}}[i] + \mathbf{E}_{\text{slot}}[r], \tag{5}$$

where $p_i^{(r)}$ is the quantized prosody value, $\mathbf{E}_{\text{wpos}}[i]$ encodes its word-level position embedding, and $\mathbf{E}_{\text{slot}}[r]$ identifies its prosodic slot based position embedding.

Because prosody depends on both current-word semantics and prior prosody, we align prosodic information with the existing word representations rather than serializing it as additional sequence tokens. This keeps the effective context length identical to that of the original text sequence, preserves the backbone's native positional structure, and avoids the roughly sevenfold sequence expansion required when each word is interleaved with six prosody tokens (Qian et al., 2025).

Specifically, we condition each word representation on its causally available prosody context. When predicting attribute r of word i, the model aggregates the complete prosody bundle of the preceding word and the already generated attributes of the current word into a context vector $\mathbf{c}_{i,r}$. We inject this context at the final-subword position $j(i)$ of word i. Let $x_{j(i)} \in \mathcal{V}_T$ denote the text token at this

position. We directly align it with the semantic representation of the complete word:

$$\widetilde{\mathbf{x}}_{j(i)}^{(r)} = \mathbf{E}_T\big[x_{j(i)}\big] + \mathbf{c}_{i,r}. \tag{6}$$

Here, $\widetilde{\mathbf{x}}_{j(i)}^{(r)}$ is the conditioned word representation used to predict attribute r. The resulting prosody-aware representations of preceding words remain accessible through causal self-attention, enabling longer-range prosodic dependencies without increasing the sequence length. The exact causal factorization and normalization of $\mathbf{c}_{i,r}$ are provided in Appendix A.2.1.

Prosody generation adapter. To preserve the pretrained linguistic and multimodal representations, we keep the backbone frozen and introduce a lightweight prosody LoRA (Hu et al., 2021) throughout the language decoder to adapt the backbone to a new numerical output space. Given the prosody-conditioned word representation $\widetilde{\mathbf{x}}_{j(i)}^{(r)}$, the adapter transforms it through the frozen backbone to predict the next quantized prosody value. For each adapted projection in decoder layer ℓ, we compute

$$\mathbf{y}^{(\ell)} = \mathbf{W}^{(\ell)}\mathbf{h}^{(\ell)} + \frac{\alpha}{\rho}\mathbf{U}_P^{(\ell)}\mathbf{D}_P^{(\ell)}\mathbf{h}^{(\ell)}, \tag{7}$$

where $\mathbf{W}^{(\ell)}$ is the frozen pretrained projection, ρ is the LoRA rank, α is its scaling factor, and $\mathbf{D}_P^{(\ell)}$ and $\mathbf{U}_P^{(\ell)}$ are trainable low-rank parameters, and $\mathbf{h}^{(\ell)}$ and $\mathbf{y}^{(\ell)}$ denote the input and output representations of the adapted projection in decoder layer ℓ, respectively. The attributes of each word are generated autoregressively in a predefined within-word order. After predicting an attribute, its quantized value is embedded and incorporated into the causal context used to predict the next attribute. Once the complete prosody bundle of word i has been generated, it conditions the prediction for word $i+1$. An output head maps the adapted word representation to the numerical vocabulary $\mathcal{V}_P$. The complete decoding formulation is provided in Appendix A.2.2.

4.3 Training and Inference

Two-stage training. Learning multimodal prosody generation from limited aligned data is challenging. We therefore adopt a two-stage training strategy that first acquires general linguistic-prosodic priors and subsequently aligns them with multimodal events.

In the first stage, we perform knowledge distillation from ProsodyLM (Qian et al., 2025). Specifically, we use the default ProsodyLM model to annotate the text, assigning word-level prosodic attributes to each token. Using these pseudo-labels, we train the prosody embedding tables, structural embeddings, prediction head, and prosody-generation LoRA while keeping the pretrained backbone frozen. This stage equips the model with a general prior over the relationship between language and prosody. In the second stage, we fine-tune the model on aligned multimodal commentary data. The objective is to associate prosody with relevant visual and acoustic events while preserving the linguistic-prosodic prior acquired during distillation. Prosodic attributes are predicted autoregressively, and all newly introduced prosody modules remain trainable. Since sports footage contains mostly routine moments, we also develop a filtering strategy to identify highly exciting scenes and construct a more balanced dataset for further fine-tuning, improving the immersiveness of commentary during key events. Further details on the distillation data and multimodal fine-tuning procedure are provided in Appendix A.2.4.

Training objective. Cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ can distinguish different bins but cannot capture the numerical ordering of quantized prosody bins. For example, for the energy attribute, bin 0 represents much lower energy than bin $Q-1$, yet cross-entropy does not explicitly account for the distance between bins. We therefore introduce a distance loss $\mathcal{L}_{\mathrm{RMSE}}$ that penalizes the discrepancy between the ground-truth value and the expected value of the predicted distribution. The overall objective is

$$\mathcal{L} = \lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{RMSE}}\mathcal{L}_{\mathrm{RMSE}}, \tag{8}$$

where λ_{CE} and λ_{RMSE} are tunable weights that balance the two loss terms to prevent either from dominating training. Detailed loss formulations are provided in Appendix A.2.5.

5 EXPERIMENT

We compare our model against several baselines and demonstrate improved alignment with ground-truth prosody and higher prosodic quality. A user study further confirms better alignment between the generated prosody and environmental cues. Ablation studies highlight the contributions of the proposed architecture and input modalities.

5.1 EXPERIMENT SETTINGS

Experiment setup. We evaluate on the held-out INF-STREAM-PROSODY test set described in Sec. 3.2, together with FIFA2026, an out-of-domain football-commentary benchmark constructed using the same preprocessing pipeline. We additionally report results on INF-STREAM-EXCITE, an expressive subset consisting of the top 30% of test samples ranked by ground-truth prosodic intensity. This subset emphasizes moments in which event-conditioned prosody is particularly important. We report two versions of our models PROSODYMLM and PROSODYMLM-BAL result. The former one is trained on the full INF-STREAM-PROSODY, and the latter one is further tuned on a balanced subset of INF-STREAM-PROSODY.

To isolate prosody generation from differences in textual content, all systems are conditioned on the same ground-truth commentary text. Unless otherwise specified, prosody is generated autoregressively. For methods that predict discrete prosody, we synthesize speech using the same StyleTTS2 decoder. To minimize the influence of background noise in the original recordings, we evaluate all methods against reference speech reconstructed from the ground-truth prosody tokens using the same decoder. Full training and implementation details are provided in Appendix A.4.1.

Baselines. We compare ProsodyMLM with the following baselines:

QWEN3-TTS: We use QWEN3-TTS (Hu et al., 2026) as a strong text-to-speech baseline. We provide the ground-truth commentary text as input and synthesize the corresponding speech. We additionally provide a sports-commentary-oriented prompt and a reference commentary utterance to encourage a timbre similar to that of the StyleTTS2-generated speech used by ProsodyLM.

PROSODYLM: We use PROSODYLM (Qian et al., 2025) as our text-conditioned prosody generation baseline. Given the ground-truth commentary text, ProsodyLM predicts discrete word-level prosody tokens, which are subsequently converted into speech.

Metrics. Following ProsodyLM (Qian et al., 2025), we evaluate how closely the generated speech matches the reference prosody along four dimensions: pitch, speaking rate, energy, and emotion. We report utterance-level absolute differences in mean F0 ($|\Delta\mathrm{F0}|$) , speaking rate ($|\Delta\mathrm{Rate}|$), and RMS-normalized Mel energy ($|\Delta\mathrm{Energy}|$), where lower values indicate better alignment. We additionally measure emotional alignment using the cosine similarity between Emotion2Vec+ (Ma et al., 2024) embeddings of the generated and reference speech (**Emo.**), where higher is better. Detailed metric definitions are provided in Appendix A.4.2.

5.2 QUANTITATIVE RESULTS

Quantitative evaluation on Inf-stream-Prosody and FIFA2026. As shown in Table 1, despite a limited training budget, our method achieves strong performance on the INF-STREAM-PROSODY benchmark, with better alignment in pitch and speaking rate, as well as higher emotional similarity to the reference speech. For energy, our method is slightly worse than ProsodyLM but remains competitive overall. These results indicate that our lightweight adaptation strategy can effectively adapt a pretrained streaming VLM for prosody generation while preserving its pretrained capabilities.

The advantage of multimodal grounding becomes more evident on the high-excitement subset, INF-STREAM-EXCITE, where appropriate delivery depends more strongly on the unfolding audio-visual events. Our model maintains strong overall performance, and finetuning on the balanced subset further improves its performance on highly expressive moments. This is particularly important because most sports footage consists of relatively routine events, while highly expressive moments occur much less frequently in the full training distribution.

Table 1: **Quantitative comparison on Inf-Stream.** We report utterance-level evaluation on INF-STREAM-PROSODY and the high-excitement INF-STREAM-EXCITE subset.

	Inf-Stream-Prosody				Inf-Stream-Excite			
System	$\lvert\Delta F0\rvert \downarrow$	$\lvert\Delta Rate\rvert \downarrow$	$\lvert\Delta En.\rvert \downarrow$	Emo. $\uparrow$	$\lvert\Delta F0\rvert \downarrow$	$\lvert\Delta Rate\rvert \downarrow$	$\lvert\Delta En.\rvert \downarrow$	Emo. $\uparrow$
Qwen3-TTS	34.29	2.163	1.211	0.939	54.49	2.179	1.209	0.926
ProsodyLM	64.42	2.627	**0.552**	0.941	46.09	2.404	**0.538**	0.947
Ours	**32.30**	2.135	0.643	**0.951**	44.81	2.246	0.677	0.948
Ours-bal	46.06	**1.958**	0.602	0.948	**42.92**	**1.895**	0.614	**0.950**

We also observe different biases in the text-conditioned baselines. Qwen3-TTS tends to produce relatively flat delivery, allowing it to perform reasonably well on the full test set, which contains mostly routine moments, but its performance degrades substantially on the high-excitement subset. Conversely, ProsodyLM tends to generate more expressive prosody, leading to relatively stronger performance on the excitement subset but weaker performance on the full benchmark. These results suggest that text-conditioned speech models have difficulty determining the appropriate degree of expressiveness for a particular event without access to its perceptual context. In comparison, our model more consistently adapts its prosody across both routine and highly expressive moments, demonstrating the benefit of grounding prosody in vision and audio.

We further evaluate generalization on FIFA2026, an out-of-distribution football-commentary benchmark, as shown in Table 2. Despite the distribution shift, our model remains competitive with the baselines across metrics and achieves stronger performance on several of them. Overall, these results suggest that our approach enables effective multimodal prosody adaptation with limited training resources and generalizes beyond the training distribution.

User study. We conduct a user study to evaluate the perceptual quality of the generated narration. Participants are presented with the reference video as well as the ground-truth narration. For each trial, participants are presented with a pair of anonymized outputs, one from our method and one from a baseline, and are asked which output better matches the reference narration. Ties are allowed when the two outputs are judged comparable. As shown in Fig. 3, our method is preferred over Qwen3-TTS in 43.6% of comparisons, compared with 30.9% for Qwen3-TTS, with 25.5% ties. The margin is even larger when compared against ProsodyLM, where our method is preferred in 57.3% of comparisons versus 19.1% for ProsodyLM, with 23.6% ties. These results complement the objective metrics and suggest that the improved prosodic alignment is also perceived by the human listeners. The details of conducting user study can be found in Appendix A.8.

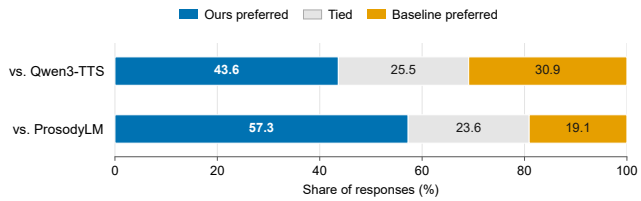

Figure 3: **User study results.** Human listeners also agree that in most cases our model aligns better with the reference than the baselines.

Ablation study. We conduct an ablation study to assess the contributions of each modality and component in our method: (1) visual input, (2) environmental sound input, (3) distillation stage, (4) the RMSE loss, and (5) prosody injection. All models are trained on the balanced subset of INF-STREAM-PROSODY and test on a randomly selected 1000 subset of the test set. The results, summarized in Table 3, confirm the effectiveness of each component. Removing either vision or audio degrades overall performance, showing that both modalities provide complementary cues for event-aligned prosody. Prosody distillation provides a strong text–prosody prior before multimodal finetuning, while the RMSE loss improves the modeling of continuous prosodic variation. Most notably, replacing our prosody-injection mechanism with serialized prosody tokens leads to a substantial performance drop, highlighting the importance of preserving the pretrained backbone's original text sequence structure. See Appendix A.6 for more details and analysis.

5.3 QUALITATIVE RESULTS

Fig. 4 presents qualitative examples of our model's narration together with the corresponding visual and acoustic context. For each example, we show the input video frames and environmental-audio

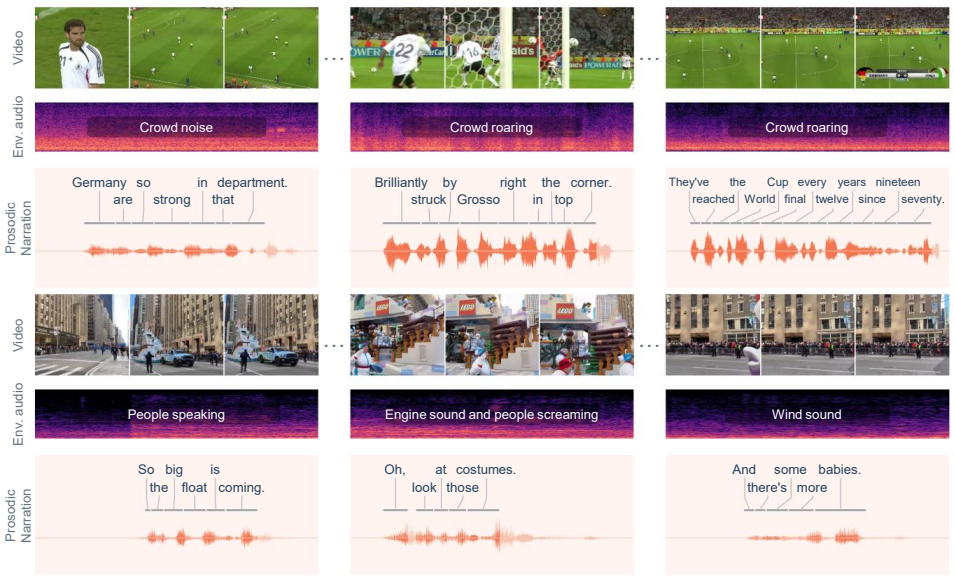

Figure 4: **Qualitative results of ProsodyMLM.** We illustrate how our model adapts narration prosody to different audio-visual contexts. In soccer videos, routine passages have relatively neutral delivery, while salient moments with excited crowd reactions produce more expressive prosody. When applied zero-shot to out-of-distribution parade videos, the model similarly becomes more excited in response to crowd screaming, suggesting that it can associate shared audio-visual cues with appropriate prosodic changes beyond sports.

Table 2: **Results on FIFA2026.** We evaluate each model's out-of-distribution generalization.

System	$\lvert\Delta\text{F0}\rvert \downarrow$	$\lvert\Delta\text{Rate}\rvert \downarrow$	$\lvert\Delta\text{En.}\rvert \downarrow$	Emo. ↑
Qwen3-TTS	111.63	3.263	1.915	0.862
ProsodyLM	96.12	2.344	1.031	0.893
Ours	**78.04**	2.650	**0.777**	**0.911**
Ours-bal	79.78	**2.287**	0.870	0.868

Table 3: **Ablation study.** We evaluate each modality and model component's contribution.

System	$\lvert\Delta\text{F0}\rvert \downarrow$	$\lvert\Delta\text{Rate}\rvert \downarrow$	$\lvert\Delta\text{En.}\rvert \downarrow$	Emo. ↑
Full	**31.47**	**2.361**	0.678	**0.944**
w/o Vision	36.46	2.417	**0.651**	0.942
w/o Audio	34.20	2.600	0.689	0.937
w/o Distillation	36.98	3.181	0.842	0.917
w/o RMSE Loss	32.64	3.090	0.865	0.935
w/o Prosody Injection	91.98	10.99	1.454	0.870

spectrogram, along with the generated speech waveform and its text transcript. The expressiveness of the generated narration is visually reflected by its waveform dynamics and speaking pace. In the soccer examples, routine passages of play are accompanied by relatively neutral delivery, whereas a scoring moment with strong crowd reactions elicits substantially more energetic prosody. These examples illustrate how the model adapts its delivery to the evolving audio-visual context. We further evaluate an out-of-distribution parade video, where the narration becomes more expressive as participants in distinctive costumes appear alongside crowd cheers. Since these acoustic cues resemble those observed in sports broadcasts, the example suggests that the model can transfer learned prosodic responses to scenarios beyond sports commentary. Additional qualitative examples are provided on the demo page and in Appendix A.5.

6 Conclusion

We introduced the *Prosodic Video Narration*, a new task of continuously generating spoken commentary with prosody grounded in streaming audio-visual events, and presented ProsodyMLM, a streaming multimodal language model for this task. By conditioning word-level prosody on streaming visual and environmental audio cues together with historical context, ProsodyMLM produces narration whose delivery better aligns with unfolding events while maintaining coherent prosody over time. Evaluations show improvements over baselines on both objective metrics and a user study. Future work will extend the approach to more diverse streaming scenarios and jointly train text and prosody generation, allowing prosodic representations to also guide what the model says.

7 AI DISCLOSURE

In this work, we used generative AI tools to improve the clarity, grammar, and readability of the manuscript, and to assist with code writing and debugging during implementation.

REFERENCES

Md Mubtasim Ahasan, Md Fahim, Tasnim Mohiuddin, AKM Rahman, Aman Chadha, Tariq Iqbal, M Ashraful Amin, Md Mofijul Islam, and Amin Ahsan Ali. Dm-codec: Distilling multimodal representations for speech tokenization. *arXiv preprint arXiv:2410.15017*, 2024.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *CVPR*, 2024.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video llm with streaming speech transcription at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29083–29095, June 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9921–9931, October 2023.

FIFA. FIFA World Cup 2026: Memorable Moments and Match Highlights. YouTube playlists, 2026. https://www.youtube.com/playlist?list=PLBRLtDhTHh5o, https://www.youtube.com/playlist?list=PLB5dkLRMn9so, accessed 2026-09-26.

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1711–1721, June 2018.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. Recent advances in discrete speech tokens: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19594–19621. Curran Associates, Inc., 2023. doi: 10.52202/075280-0860. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3eaad2a0b62b5ed7a2e66c2188bb1449-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *Proc. ACL 2024 Findings*, 2024.

Kaizhi Qian, Xulin Fan, Junrui Ni, Slava Shechtman, Mark A. Hasegawa-Johnson, Chuang Gan, and Yang Zhang. ProsodyLM: Uncovering the emerging prosody processing capabilities in speech language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=uBg8PClMUu.

Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. MatchTime: Towards automatic soccer game commentary generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1671–1685, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.99. URL https://aclanthology.org/2024.emnlp-main.99/.

Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards universal soccer video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8384–8394, June 2025.

Amélie Royer, Moritz Böhle, Laurent Mazaré, Neil Zeghidour, Alexandre Défossez, and Patrick Pérez. Vision-speech models: Teaching speech models to converse about images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1706–1715, 2026.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast, 2025. URL https://arxiv.org/abs/2509.14128.

Slava Shechtman and Raul Fernandez. A neural tts system with parallel prosody transfer from unseen speakers. In *Interspeech*, 2023. URL https://api.semanticscholar.org/CorpusID:260907834.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer, Piotr Dollár, Wei-Ning Hsu, and Ann Lee. Sam audio: Segment anything in audio, 2025. URL https://arxiv.org/abs/2512.18099.

Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir Mohammad Mansourian, Xin Zhou, Shohreh Kasaei, Bernard Ghanem, Alexandre Alahi, Marc Van Droogenbroeck, and Christophe De Vleeschouwer. Soccernet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3293–3305, 2024. doi: 10.1109/CVPRW63382.2024.00334.

Kim Sung-Bin, Jeongsoo Choi, Puyuan Peng, Joon Son Chung, Tae-Hyun Oh, and David Harwath. Voicecraft-dub: Automated video dubbing with neural codec language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14623–14632. IEEE, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James

O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL `https://arxiv.org/abs/2607.02770`.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Haotian Xia, Haonan Ge, Junbo Zou, Hyun Woo Choi, Xuebin Zhang, Danny Suradja, Botao Rui, Ethan Tran, Wendy Jin, Zhen Ye, Xiyang Lin, Christopher Lai, Shengjie Zhang, Junwen Miao, Shichao Chen, Rhys Tracy, Vicente Ordonez, Weining Shen, and Hanjie Chen. Sportr: A benchmark for multimodal large language model reasoning in sports. In *International Conference on Learning Representations (ICLR)*, 2026.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Jinglin Xu, Guohao Zhao, Sibo Yin, Wenhao Zhou, and Yuxin Peng. Finesports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21773–21782, June 2024.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. In *International Conference on Learning Representations*, volume 2026, pp. 61463–61475, 2026.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, 2023.

A APPENDIX

A.1 DATASET

Audio-commentary separation. Given a sports video, we first separate it into visual and audio streams. We then use the text-guided audio segmentation model `sam-audio-base` (Shi et al., 2025) to disentangle spoken commentary from environmental sound. Specifically, we use the prompt *"human speech, male voice, female voice, speaking, talking, announcer"* to extract the commentary and treat the residual signal as environmental audio. This produces temporally aligned commentary and environmental-audio streams while retaining their mixture as the original full-audio stream.

ASR and discrete attributes captioning. After separating the spoken commentary from the environmental audio, we first transcribe the commentary using the NVIDIA Parakeet-TDT ASR model, `nvidia/parakeet-tdt-0.6b-v2` (Sekoyan et al., 2025). We retain the word-level timestamps predicted by the ASR model and use them to associate each lexical item with its corresponding speech region. We then follow the preprocessing procedure of ProsodyLM (Qian et al., 2025) to obtain a word-aligned discrete prosody representation.

Specifically, each word is represented by one preceding-silence value and five prosodic attributes, denoted as $P1$–$P5$. $P1$ characterizes word duration and is computed from the mean soft-aligned phoneme duration in the logarithmic domain. $P2$ captures the pitch range within the word, measured as the difference between the 95th and 5th percentiles of voiced log-F0 values. $P3$ represents the overall pitch level of the word and is defined as the median voiced log-F0. $P4$ describes the local pitch dynamics by fitting a least-squares slope to the voiced log-F0 trajectory over Mel-frame indices. Finally, $P5$ represents word-level energy using the log-energy statistic extracted by the StyleTTS2-based acoustic frontend. Frames with estimated F0 below 50 Hz are treated as unvoiced when computing the pitch-related attributes.

The continuous value of each attribute is converted into a discrete numerical representation using the fixed normalization and quantization scheme adopted from ProsodyLM (Qian et al., 2025). For each valid attribute value, we first normalize it to a predefined range using attribute-specific statistics and then uniformly quantize the normalized value into 512 numerical bins, indexed from 0 to 511. We additionally reserve index 512 as a sentinel class for missing or invalid measurements. Therefore, the resulting numerical prediction space contains

$$Q = 512 + 1 = 513$$

classes in total. The sentinel should be interpreted as a missing-value indicator rather than an additional uniformly spaced measurement bin.

For each word, the resulting structured prosody representation is therefore composed of

$$\left(p_{\text{sil}}, p^{(1)}, p^{(2)}, p^{(3)}, p^{(4)}, p^{(5)}\right),$$

where p_{sil} denotes the preceding-silence value and $p^{(1)}$–$p^{(5)}$ correspond to the five quantized prosodic attributes described above. This word-level representation is used as the supervision target for both prosody distillation and multimodal finetuning.

FIFA 2026 The FIFA2026 dataset was created by joining a set of memorable moments and highlights with selected plain moments of the 2026 World Cup. Memorable moments and match highlights were collected from two publicly available YouTube playlists (FIFA, 2026), while plain moments were selected manually from full game replays at points where no notable event occurs. The dataset consists of 206 memorable moments, 102 highlights, and 100 plain clips.

A.2 METHOD DETAILS

A.2.1 WORD-LEVEL PROSODY INJECTION

Separate numerical vocabulary. Let $\mathcal{V}_T$ denote the pretrained text vocabulary and $\mathcal{V}_P = \{0, \ldots, Q-1\}$ a disjoint numerical vocabulary, where Q is the number of quantization bins defined in Sec. 3.1. The pretrained text embedding table $\mathbf{E}_T$ remains frozen. We introduce a trainable

prosody embedding table

$$\mathbf{E}_P \in \mathbb{R}^{Q\times d}, \tag{9}$$

where d is the hidden dimension of the backbone. Since prosodic attributes are correlated and share the same quantized value range, they use the same vocabulary $\mathcal{V}_P$ and embedding table $\mathbf{E}_P$. This sharing reduces the number of trainable parameters and provides a compact prosody prediction space. In this project, we set $Q = 513$, which contains 512 ordinary quantization bins and one missing-value sentinel. As we are utilizing the `mit-han-lab/StreamingVLM` as the VLM backbone, our $d = 3584$.

Word-attribute association. A quantized value does not identify its associated word or prosodic role by itself. We therefore introduce a trainable prosody word-index embedding $\mathbf{E}_{\text{wpos}}$ and a trainable attribute-slot embedding $\mathbf{E}_{\text{slot}}$. For attribute $r \in \mathcal{R}$ of word i, its side-channel representation is

$$\mathbf{z}_{i,r} = \mathbf{E}_P[p_i^{(r)}] + \mathbf{E}_{\text{wpos}}[i] + \mathbf{E}_{\text{slot}}[r], \tag{10}$$

where $p_i^{(r)} \in \mathcal{V}_P$ and $r \in \mathcal{R}$, with $\mathcal{R}$ defined in Sec. 3.1. All three prosody embedding tables are newly introduced and optimized during training; they are distinct from the frozen text embedding table $\mathbf{E}_T$.

Causal prosody context. Let $\prec$ denote the predefined attribute order within the word. When predicting attribute $r^\star$ of word i, define

$$\mathcal{R}_{<r^\star} = \{r \in \mathcal{R} : r \prec r^\star\}. \tag{11}$$

The available prosody context is

$$\mathbf{c}_{i,r^\star} = \frac{1}{\sqrt{|\mathcal{R}|}} \left(\mathbb{I}[i > 0] \sum_{r\in\mathcal{R}} \mathbf{z}_{i-1,r} + \sum_{r\in\mathcal{R}_{<r^\star}} \mathbf{z}_{i,r} \right). \tag{12}$$

The indicator $\mathbb{I}[i > 0]$ removes the preceding-word term for the first word. Thus, the model observes the complete prosody bundle of the immediately preceding word and only the already generated attributes of the current word. The factor $1/\sqrt{|\mathcal{R}|}$ is a fixed normalization based on the maximum size of a complete attribute bundle. It limits the magnitude of the summed residual and prevents the side channel from dominating the native text representation. As there are 6 different attribute tokens, $|\mathcal{R}| = 6$.

Injection and positional encoding. Let $j(i)$ denote the position of the final subword token of word i in the original serialized multimodal sequence, and let $x_{j(i)} \in \mathcal{V}_T$ denote the text token at this position. Its native input embedding is $E_T[x_{j(i)}]$. We inject the causal prosody context directly into this token representation:

$$\tilde{x}_{j(i)}^{(r^\star)} = E_T[x_{j(i)}] + c_{i,r^\star}. \tag{13}$$

By injecting the prosodic attributes into the representation of the corresponding word, rather than inserting additional prosody tokens into the serialized sequence, we preserve the original token layout and relative positions of all text tokens. Consequently, the backbone can retain its native RoPE positional encoding without modification, allowing its pretrained language understanding and generation capabilities to be reused directly. In this formulation, the backbone's RoPE continues to encode each token's position in the serialized multimodal sequence, while $\mathbf{E}_{\text{wpos}}$ and $\mathbf{E}_{\text{slot}}$ separately encode the internal word–attribute structure of the prosody side channel.

A.2.2 AUTOREGRESSIVE DECODING

Starting from $\tilde{\mathbf{x}}_{j(i)}^{(r)}$, the adapted backbone produces a contextualized hidden state $\mathbf{h}_{j(i)}^{(r)}$. A shared prosody prediction head then maps this state to logits over the numerical vocabulary:

$$\mathbf{o}_{i,r} = \mathbf{W}_P \mathbf{h}_{j(i)}^{(r)} + \mathbf{b}_P, \qquad \mathbf{o}_{i,r} \in \mathbb{R}^Q, \tag{14}$$

where $\mathbf{W}_P$ and $\mathbf{b}_P$ are trainable and shared across all prosodic attributes. Let $\mathbf{x}$ denote the commentary token sequence, $\mathbf{p}_{<i}$ the prosody bundles of all preceding words, and $\mathbf{p}_i^{(<r)}$ the already

generated prosodic attributes of word i that precede attribute r in the predefined within-word order. The distribution of the next prosody value is

$$P\left(p_i^{(r)} = q \mid \mathbf{x}, \mathbf{p}_{<i}, \mathbf{p}_i^{(<r)}\right) = \mathrm{softmax}\left(\mathbf{o}_{i,r}\right)_q, \qquad q \in \mathcal{V}_P. \tag{15}$$

At inference time, the next quantized value is selected from this distribution:

$$\widehat{p}_i^{(r)} = \arg\max_{q \in \mathcal{V}_P} \mathbf{o}_{i,r}[q]. \tag{16}$$

The predicted value is embedded using $\mathbf{E}_P$ and added to the causal prosody context before predicting the next attribute. Repeating this procedure according to the predefined within-word attribute order yields the complete prosody bundle for word i, which is then used to condition the prediction of word $i+1$.

A.2.3 PROMPT DETAIL

We train and inference our model using the following details. During the training we only calculate the prosody attributes in the target commentary and the prosody tokens are injected using the formula above.

Previous vision-audio content:
`<mean-pooled previous vision>`
`<mean-pooled previous audio>`

Current vision-audio content:
`<current vision tokens>`
`<current audio tokens>`

Please generate live sports commentary with natural speech prosody using the provided audio-video context.

Assistant:
`<masked historical commentary with SIL and P1--P5> <target commentary with SIL and P1--P5>`

A.2.4 TRAINING STRATEGY DETAILS

Trainable components. We adopt StreamingVLM (Xu et al., 2026) as the multimodal backbone, leveraging its pretrained visual encoder and language model for video understanding and commentary modeling. To incorporate acoustic context, we use a pretrained Wav2Vec 2.0 encoder (Baevski et al., 2020) to extract frame-level audio features, which are mapped into the language-model hidden space through a trainable two-layer MLP with GELU activations.

Throughout training, the visual encoder, visual projector, language-model backbone, and audio encoder remain frozen. We only optimize the newly introduced prosody-related components, including the numerical prosody embeddings, word-index and attribute-slot embeddings, audio projector, prosody LoRA modules, and numerical prediction heads. This lightweight adaptation preserves the pretrained multimodal and linguistic knowledge of the backbone while introducing the capacity required for word-level prosody prediction.

Stage I: prosody distillation. We first initialize the prosody modules by distilling general text–prosody knowledge from ProsodyLM (Qian et al., 2025). Given a commentary sentence $\mathbf{W_s}$, the frozen teacher predicts word-aligned pseudo prosody labels

$$\widetilde{\mathbf{P}} = f_{\mathrm{ProsodyLM}}(\mathbf{W_s}). \tag{17}$$

The student is then trained to predict $\widetilde{\mathbf{P}}$ conditioned on the commentary text. This stage provides the newly introduced prosody modules with a general prior over the relationships between semantics, linguistic context, and word-level prosody before multimodal conditioning is introduced. Specifically, we randomly sample 1M captions from LiveCC-5M (Chen et al., 2025) and generate their pseudo prosody labels using ProsodyLM. Since LiveCC-5M covers diverse video domains beyond sports, this distillation stage also provides a broader prosodic prior rather than specializing the model prematurely to sports commentary.

Stage II: multimodal finetuning. Starting from the distilled checkpoint, we optimize all trainable prosody-related components on INF-STREAM-PROSODY. The model is conditioned on the streaming visual and environmental-audio context together with the commentary text, allowing the text-derived prosody prior learned in Stage I to adapt to temporally aligned audio-visual events in live sports broadcasts.

P3-based event-aware sampling. Sports broadcasts are dominated by routine segments with relatively neutral delivery, while salient events associated with expressive prosody occur much less frequently. To increase the exposure to such events during training, we use P3, which represents median F0, as a simple proxy for prosodic intensity. Generally, a high F0 leads to higher excitement level while delivering. For each utterance, we compute the mean P3 value across its words and rank all training samples accordingly. We retain the top 20% highest-P3 utterances and randomly sample another 20% from the remaining data. The latter preserves a diverse set of ordinary commentary examples and avoids reducing the training distribution to an artificial separation between highly expressive and strictly neutral speech. We denote the resulting subset as INF-STREAM-PROSODY-BAL. After finetuning on the full INF-STREAM-PROSODY training set, we further finetune the model on this event-aware balanced subset to improve its sensitivity to salient and expressive moments.

A.2.5 PROSODY TRAINING OBJECTIVE

For each prosodic attribute $r \in \mathcal{R}$, let Ω_r denote the set of valid word positions for attribute r. Let $\mathbf{x}$ denote the commentary token sequence and $\mathbf{C}$ the current and historical multimodal context. For $n \in \Omega_r$, the model produces logits $\mathbf{o}_{n,r} \in \mathbb{R}^Q$ and

$$\pi_{n,r}(v) = P_\theta\left(p_n^{(r)} = v \mid \mathbf{x}, \mathbf{p}_{<n}, \mathbf{p}_n^{(<r)}, \mathbf{C}\right) = \mathrm{softmax}(\mathbf{o}_{n,r})_v. \tag{18}$$

For each attribute, we compute the categorical cross-entropy

$$\mathcal{L}_{\mathrm{CE}}^{(r)} = -\frac{1}{|\Omega_r|} \sum_{n \in \Omega_r} \log \pi_{n,r}\left(p_n^{(r)*}\right), \tag{19}$$

where $p_n^{(r)*}$ is the ground-truth quantized value.

Although cross-entropy distinguishes the quantization bins, it does not explicitly encode their ordinal relationships. We therefore introduce a distance loss. We first compute the expected numerical value over the ordinary quantization bins, excluding the missing-value sentinel:

$$\mu_{n,r} = \sum_{v=0}^{Q-2} v\, \pi_{n,r}(v). \tag{20}$$

Let $\Omega_r^{\mathrm{num}} \subseteq \Omega_r$ denote the positions whose ground-truth targets correspond to ordinary quantization bins rather than the missing-value sentinel. We define

$$\mathcal{L}_{\mathrm{RMSE}}^{(r)} = \sqrt{\frac{1}{|\Omega_r^{\mathrm{num}}|} \sum_{n \in \Omega_r^{\mathrm{num}}} \left(\mu_{n,r} - p_n^{(r)*}\right)^2}. \tag{21}$$

The losses are averaged across the attributes present in the batch:

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{|\mathcal{R}'|} \sum_{r \in \mathcal{R}'} \mathcal{L}_{\mathrm{CE}}^{(r)}, \qquad \mathcal{L}_{\mathrm{RMSE}} = \frac{1}{|\mathcal{R}'|} \sum_{r \in \mathcal{R}'} \mathcal{L}_{\mathrm{RMSE}}^{(r)}, \tag{22}$$

where $\mathcal{R}' \subseteq \mathcal{R}$ contains the attributes with valid targets in the current batch. The final training objective is

$$\mathcal{L} = \lambda_{\mathrm{CE}} \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{RMSE}} \mathcal{L}_{\mathrm{RMSE}}. \tag{23}$$

During training, we optimize the prosody-related parameters together with the audio projector, while keeping the pretrained backbone frozen. We chose $\lambda_{\mathrm{CE}} = 1$ and $\lambda_{\mathrm{RMSE}} = 0.1$, which make both losses in the same scale.

A.3 INFERENCE

Given the commentary text and its multimodal context, the model generates the prosodic attributes of each word autoregressively in a predefined order. Each predicted value is embedded and added to the causal context used for the subsequent attribute, and the completed bundle conditions the next word. The visual features are encoded once and shared across all attribute predictions, avoiding repeated visual encoding. This procedure preserves the original text sequence length and produces a word-aligned prosody sequence without interleaving additional prosody tokens with the text.

A.4 IMPLEMENTATION DETAILS

A.4.1 EXPERIMENTAL SETTING

We train all models using four NVIDIA A6000 GPUs. During the distillation stage, we train for 200K steps with a per-GPU batch size of 2 and a context length of 1,024. We use AdamW with a weight decay of 0.01, a learning rate of 1×10^{-4}, cosine learning-rate decay, and a warmup ratio of 0.03. For multimodal fine-tuning, we train PROSODYMLM on INF-STREAM-PROSODY for 100K steps with a per-GPU batch size of 1 and a context length of 1,024. We use AdamW with a weight decay of 0.01, a learning rate of 1×10^{-4}, cosine learning-rate decay, and a warmup ratio of 0.05. We retain two historical visual and audio segments and one segment of text history. For the current segment, visual inputs are sampled at 2 FPS and audio features at 10 Hz. For the prosody LoRA, we set the rank to $\rho = 16$, the scaling factor to $\alpha = 16$, and the dropout rate to 0.05, and apply LoRA to all attention and MLP projection layers. We then fine-tune the model for an additional 30K steps on the balanced subset, INF-STREAM-PROSODY-BAL, using the same training configuration.

For the baselines, we prompt Qwen3-TTS-Base (Hu et al., 2026) to act as a sports commentator and provide a StyleTTS2-generated reference utterance for timbre cloning, making the differences in prosodic delivery across methods easier to perceive. For ProsodyLM (Qian et al., 2025), we use the default inference settings provided by the authors. Given the ground-truth commentary text, Qwen3-TTS-Base directly synthesizes the corresponding speech, whereas ProsodyLM first predicts a prosodic representation, which is then used for speech synthesis. All models are evaluated on video clips ranging from 5 to 30 seconds. Clip boundaries are placed at natural pauses in the commentary whenever possible; an additional boundary is introduced when the ASR system fails to recover a sufficiently long span of valid commentary text. The statistics of the INF-STREAM-PROSODY test set and FIFA2026 are the same as those reported in Sec. 3.2. For INF-STREAM-EXCITE, we follow the same P3-based ranking procedure used to construct the balanced training subset and retain the top 30% of test utterances with the highest mean P3 values.

A.4.2 METRICS

We provide detailed definitions of the objective metrics used to evaluate prosodic alignment between generated and reference speech. Adapting from ProsodyLM (Qian et al., 2025), all metrics are computed at the utterance level and then averaged across the evaluation set. Since the generated and reference speech share the same commentary text, these metrics primarily capture differences in speech delivery rather than linguistic content.

Pitch Difference ($|\Delta\mathrm{F0}|$). We extract the F0 contour of each utterance and compute the mean F0 over voiced frames. For each generated–reference pair, we report the absolute difference between their mean F0 values:

$$|\Delta\mathrm{F0}| = \left|\overline{\mathrm{F0}}_{\mathrm{gen}} - \overline{\mathrm{F0}}_{\mathrm{ref}}\right|. \tag{24}$$

This metric measures utterance-level pitch alignment between the generated and reference speech. Lower values indicate closer agreement with the reference pitch.

Speaking-Rate Difference ($|\Delta\mathrm{Rate}|$). Let N_{phoneme} denote the number of phoneme symbols in the commentary and D the duration of the corresponding utterance. We define the speaking rate as

$$\mathrm{Rate} = \frac{N_{\mathrm{phoneme}}}{D}. \tag{25}$$

Because the generated and reference speech contain the same text, N_{phoneme} is identical for each pair, and differences in speaking rate therefore reflect differences in duration and delivery speed. We

report

$$|\Delta\mathrm{Rate}| = |\mathrm{Rate}_{\mathrm{gen}} - \mathrm{Rate}_{\mathrm{ref}}|. \tag{26}$$

Lower values indicate better alignment with the reference speaking pace.

Energy Difference ($|\Delta\mathrm{Energy}|$). To measure differences in energy patterns while reducing the effect of global loudness mismatch, we first perform paired-reference RMS normalization for each generated utterance. All waveforms are converted to mono and resampled to 24 kHz. Given a generated waveform x_{pred} and its paired reference x_{gt}, we compute their utterance-level RMS values as

$$r_{\mathrm{gt}} = \sqrt{\mathrm{mean}(x_{\mathrm{gt}}^2)}, \qquad r_{\mathrm{pred}} = \sqrt{\mathrm{mean}(x_{\mathrm{pred}}^2)}. \tag{27}$$

The generated waveform is then normalized using the gain

$$g = \frac{r_{\mathrm{gt}}}{r_{\mathrm{pred}}}, \tag{28}$$

such that its overall RMS matches that of the paired reference.

We compute an 80-bin Mel power spectrogram using a 24 kHz sampling rate, an FFT size of 2048, a window length of 1200, and a hop length of 300. Let N_{frame} denote the number of Mel frames, and let $\mathrm{Mel}(x)_t$ denote the Mel power spectrum of waveform x at frame t. Its utterance-level energy statistic is

$$E(x) = \frac{1}{N_{\mathrm{frame}}} \sum_{t=1}^{N_{\mathrm{frame}}} \log \|\mathrm{Mel}(x)_t\|_2 \,. \tag{29}$$

where $\mathrm{Mel}(x)_t$ denotes the Mel power spectrum at frame t. Since scaling the waveform by g scales the Mel power spectrogram by g^2, the normalized prediction energy can be computed as

$$E_{\mathrm{pred,norm}} = E_{\mathrm{pred,raw}} + 2\log\frac{r_{\mathrm{gt}}}{r_{\mathrm{pred}}}. \tag{30}$$

We finally report the absolute difference

$$|\Delta\mathrm{Energy}| = |E_{\mathrm{pred,norm}} - E_{\mathrm{gt}}|\,, \tag{31}$$

averaged over all utterances. Lower values indicate closer energy alignment with the paired reference after controlling for global RMS loudness.

Emotion Similarity (Emo.). We use a pretrained Emotion2Vec+ (Ma et al., 2024) model to extract an utterance-level emotion embedding from both the generated and reference speech. We measure their cosine similarity:

$$\mathrm{Emo.} = \frac{\mathbf{e}_{\mathrm{gen}}^{\top}\mathbf{e}_{\mathrm{ref}}}{\|\mathbf{e}_{\mathrm{gen}}\|_2\|\mathbf{e}_{\mathrm{ref}}\|_2}, \tag{32}$$

where $\mathbf{e}_{\mathrm{gen}}$ and $\mathbf{e}_{\mathrm{ref}}$ denote the corresponding Emotion2Vec embeddings. Higher similarity indicates that the generated speech more closely matches the emotional expression of the reference.

A.5 More qualitative results

Fig. 5 presents additional qualitative examples from two soccer and two basketball broadcasts, together with representative failure cases. Across the successful examples, the generated prosody aligns well with salient visual and acoustic events in both sports. The failure cases mainly arise when the multimodal cues become unreliable or inconsistent. For example, when the visual stream is severely blurred or the ball is barely visible, the model receives insufficient visual evidence to guide prosody generation. Failures can also occur when the visual and audio streams are temporally misaligned, causing one modality to dominate the prediction and resulting in prosody that does not fully match the underlying event.

A.6 Ablation details

We conduct an ablation study in Table 3 to evaluate the contribution of individual model components and input modalities. All variants are trained on the balanced subset used for multimodal finetuning and evaluated on 1,000 randomly selected utterances from Inf-Stream-Prosody's test set. Unless otherwise specified, all variants use the same architecture and training recipe.



Figure 5: **Additional qualitative examples.** Each column shows the input video frames, the environmental-audio spectrogram, and the generated narration waveform with its word-aligned transcript. Rows (a)–(d) contain three time-ordered clips from the same broadcast, two soccer and two basketball games. Row (e) shows failure cases from different videos.

Modality. To understand the contribution of each input modality, we train variants with either the visual or environmental-audio input removed. Removing vision substantially degrades pitch alignment, increasing the F0 difference from 31.47 to 36.46, while also slightly worsening speaking rate and emotion similarity. Removing audio similarly increases the F0 difference to 34.20 and produces a larger degradation in speaking rate and emotional similarity. These results suggest that both visual and acoustic context provide useful cues for predicting scene-aligned prosody. Visual observations help capture event dynamics, while environmental audio provides complementary cues such as crowd reactions and changes in scene intensity. Interestingly, the energy is not that kind of sensitive to the modality input, which gives a general similar result.

Distillation. To evaluate how the distillation stage helps, we remove the prosody distillation stage and train directly on the multimodal finetuning data. This causes a substantial degradation across all metrics: the F0 difference increases from 31.47 to 36.98, the speaking-rate difference from 2.361 to 3.181, and the energy difference from 0.678 to 0.842, while emotion similarity decreases from 0.944 to 0.917. These results highlight the importance of first learning a general text-prosody prior before multimodal grounding. Without this initialization, the model must learn both prosodic structure and its alignment with audio-visual events from the limited multimodal dataset, making optimization more difficult.

RMSE Loss. Removing the RMSE loss also degrades overall performance, particularly for speaking rate, whose difference increases from 2.361 to 3.090. This supports our motivation for introducing an ordinal distance objective in addition to cross-entropy: neighboring quantization bins correspond to similar prosodic values, but cross-entropy alone does not explicitly model this numerical structure. The distance-aware objective therefore provides additional supervision for learning continuous prosodic variation from discrete bins.

Prosody Injection. We train an autoregressive variant that serializes prosody as additional tokens instead of using our prosody-injection mechanism. Under the same training recipe, this variant fails to converge well. We attribute this degradation to two factors. First, interleaving text and prosody tokens substantially changes the sequence structure seen by the pretrained backbone: the original positional encoding no longer directly corresponds to the pretrained text sequence, and adapting to this new structure through LoRA is difficult with limited prosody-annotated data. In contrast, our injection mechanism preserves the backbone's original token positions while incorporating prosodic context directly into the corresponding word representations. Second, introducing prosody as additional tokens expands both the sequence length and decoding space, further increasing the optimization difficulty. Notably, this variant performs considerably worse than our model trained without distillation, suggesting that the proposed prosody injection is important for effectively adapting the pretrained backbone to prosody generation.

A.7 Additional Experiment

To examine whether directly introducing multimodal context into ProsodyLM improves prosody generation, we additionally construct an audio-visual variant, denoted as ProsodyLM-AV. We initialize the model from the publicly released ProsodyLM checkpoint and augment it with visual and environmental-audio inputs. Specifically, visual and audio features are extracted using off-the-shelf encoders (Bolya et al., 2025) and projected into the hidden space of the ProsodyLM language model through trainable MLP projectors. The resulting multimodal model is then fine-tuned on Inf-Stream-Prosody using the same aligned visual and acoustic context used by our method.

The results on FIFA2026 are reported in Table 4. Incorporating audio-visual context improves ProsodyLM on several metrics, suggesting that perceptual context provides useful guidance for prosody prediction. However, ProsodyLM-AV remains worse than our method on most metrics. One possible reason is that directly adapting the original ProsodyLM architecture to multimodal inputs may be less effective under limited multimodal supervision. In contrast, our lightweight adaptation strategy is explicitly designed to integrate streaming visual and acoustic context while preserving the pretrained backbone, leading to more consistent performance across metrics.

Table 4: **Comparison with ProsodyLM-AV on FIFA2026.** Lower is better for F0, speaking-rate, and energy differences; higher is better for emotional similarity.

System	$\lvert\Delta F0\rvert \downarrow$	$\lvert\Delta Rate\rvert \downarrow$	$\lvert\Delta En.\rvert \downarrow$	Emo. $\uparrow$
ProsodyLM	96.12	2.344	1.031	0.893
ProsodyLM-AV	94.71	2.903	**0.747**	<u>0.907</u>
Ours	**78.04**	2.650	<u>0.777</u>	**0.911**
Ours-bal	<u>79.78</u>	**2.287**	0.870	0.868

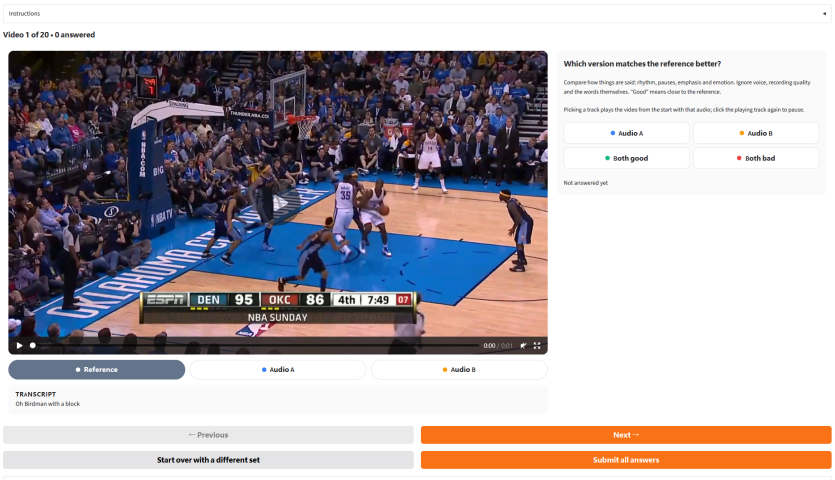

Figure 6: **User study interface.** Participants compare two anonymized generated speech samples against the reference commentary while viewing the corresponding video clip and transcript.

A.8 User study details

The listening study evaluates how faithfully different speech-generation systems reproduce the prosodic characteristics of the original commentary. Specifically, we compare ProsodyMLM with ProsodyLM and Qwen3-TTS, and assess whether the generated speech better matches the reference narration in terms of rhythm, pausing, emphasis, and emotional delivery.

Participants were presented with 20 short video clips, each accompanied by its transcript and ground-truth commentary. For each clip, we generated speech using three systems: ProsodyMLM (ours), ProsodyLM, and Qwen3-TTS. We developed a listening-study interface, shown in Fig. 6. Before beginning the study, participants were first presented with an instruction page explaining the task and evaluation criteria.

Each trial presents a video clip and its transcript together with three audio samples: the ground-truth reference commentary, a sample generated by ProsodyMLM, and a sample generated by one of the two baselines, either ProsodyLM or Qwen3-TTS. The two generated samples are anonymized as "Audio A" and "Audio B", and their presentation order is randomized to reduce ordering bias.

Participants are asked to judge which generated sample better matches the reference commentary, considering its rhythm, pauses, emphasis, and emotional delivery. They select one of four responses: "Audio A", "Audio B", "Both good", or "Both bad". In the main paper, the latter two responses are grouped together and reported as ties.