Seeing is Believing: Emotion-Aware Audio-Visual Language Modeling for Expressive Speech Generation

Weiting Tan[♠*] **Jiachen Lian**[♡] **Hirofumi Inaguma**[♡]
Paden Tomasello[♡] **Philipp Koehn**[♠] **Xutai Ma**[♡]
[♠]Johns Hopkins University [♡]Meta AI Research

Abstract

We present an Audio-Visual Language Model (AVLM) for expressive speech generation by integrating full-face visual cues into a pre-trained expressive speech model. We explore multiple visual encoders and multimodal fusion strategies during pre-training to identify the most effective integration approach. Subsequent fine-tuning on emotion recognition and expressive dialogue tasks yields substantial gains over speech-only baselines (e.g., +5 F1 in emotion recognition). AVLM highlights the value of expressive visual information in guiding speech generation and offers a foundation for end-to-end multimodal conversational systems.[1]

1 Introduction

Expressive speech generation are supported by high-quality speech tokenization/codec (Défossez et al., 2022; Jenrungrot et al., 2023; Du et al., 2024; Nguyen et al., 2024), emotionally-nuanced speech datasets (Chu et al., 2024; Huang et al., 2025; KimiTeam et al., 2025), and effective emotion-aware representation learning (Wang et al., 2023; Tang et al., 2024b). While recent advances in speech modeling have greatly improved the quality and controllability of synthesized speech, most methods rely solely on audio input and overlook the rich expressive cues available in the visual modality.

Human communication, however, is inherently multimodal. Visual signals—such as facial expressions, head gestures, and eye movements—are tightly intertwined with vocal expression, offering critical paralinguistic information that reflects emotion and intent. This raises a central question: *can we enhance expressive speech generation by incorporating these visual cues, enabling the model to produce responses that are not only fluent but also emotionally resonant and contextually adaptive?*

Figure 1: Our Audio-Visual Language Model (AVLM) perceives an audio-visual input, determine the emotion of the speaker and generate an expressive response.

To address this question, we propose developing an emotion-aware audio-visual perceptual model for expressive speech generation. Although prior work has explored lip reading, audio-visual speech recognition (AVSR), and emotion recognition (Shi et al., 2022; Hong et al., 2023; Han et al., 2024; Yeo et al., 2025a,b; Haliassos et al., 2024; Busso et al., 2008), these efforts are often task-specific. AVSR methods, for example, improve transcription in noisy settings by focusing on the lip region, but they generally ignore non-semantic yet affective visual signals. Similarly, emotion recognition benchmarks like IEMOCAP (Busso et al., 2008) are designed for classification rather than generation, limiting their utility in expressive synthesis tasks. Consequently, current systems fall short in capturing the full emotional dynamics present in audio-visual communication.

In this work, we propose an Audio-Visual Language Model (AVLM) that integrates full-face visual cues into a pre-trained expressive Speech Language Model (SpeechLM) to support emotionally rich speech generation. Our training follows a two-stage framework. First, we introduce a modality

[*] Work was done during an internship at Meta AI.

[1]All experiments, data collection, and processing activities were conducted by JHU. Meta was involved solely in an advisory role and no experiments, data collection or processing activities were conducted on Meta infrastructure. Code accessible at:https://github.com/steventan0110/AVLM

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2600–2617

fusion module to align visual representations with SpeechLM's latent space, systematically evaluating visual encoders and fusion strategies during pre-training. Second, we fine-tune the AVLM for emotion recognition and dialogue generation using synthetic data derived from IEMOCAP. To construct expressive conversations, we rewrite dialogue responses using GPT-4 (OpenAI et al., 2024) and synthesize audio with Step-Audio (Huang et al., 2025), an expressive TTS system supporting voice cloning and natural language instruction.

Our approach effectively incorporates a wider range of visual information beyond lip movements. It achieves lower perplexity during pre-training and improves AVSR performance by reducing Word Error Rate (WER) over 1 point under clean and noisy conditions. When fine-tuned for expressive speech generation and emotion recognition, our visually enhanced AVLM consistently outperforms the speech-only counterpart (by more than 5 F1-score), demonstrating higher classification accuracy and generating more expressive outputs. In summary, we strengthen the connection between visual and audio modality for expressive generation and show that our novel AVLM could be a strong foundation for building emotionally intelligent end-to-end conversational agents.

2 Related Work

2.1 Audio-Visual Speech Recognition

AVSR leverages both audio and visual modalities for robust speech transcription. Datasets like LRS3 (Afouras et al., 2018) and VoxCeleb2 (Chung et al., 2018) have driven recent progress. AV-HuBERT (Shi et al., 2022) aligns modalities via self-supervised learning, extended to multilingual settings by XLAVS-R (Han et al., 2024). Unified-Speech (Haliassos et al., 2024) incorporates auxiliary tasks to improve multi-modal alignment, while LLaMa-AVSR (Cappellazzo et al., 2025) and MMS-LLaMa (Yeo et al., 2025b) employ LLMs for improved transcription quality.

2.2 Audio-Visual Emotion Recognition

Multimodal emotion recognition has been supported by datasets such as IEMOCAP (Busso et al., 2008), RECOLA (Ringeval et al., 2013), CREMA-D (Cao et al., 2014), MSP-IMPROV (Busso et al., 2017), RAVDESS (Livingstone and Russo, 2018), and Aff-Wild (Kollias et al., 2019). Labels range from discrete classes to continuous dimensions like Valence, Arousal, and Dominance (Mehrabian and Russell, 1974). Based on these datasets, fusion methods (Savchenko, 2022; Praveen et al., 2023; Ma et al., 2024a) have been proposed to improve multimodal emotion recognition, though they are limited to classification or regression settings.

2.3 Speech Language Models

Speech Language Models (SpeechLMs) are language models trained on large text and speech datasets, typically pre-trained on text and fine-tuned with speech-text or speech-only data (Wu et al., 2023; Yu et al., 2023; Tan et al., 2024; Zhang et al., 2023; Tang et al., 2024a; Chu et al., 2024).

More recently, directly encoding speech as tokens and integrating them into pre-trained LLMs has gained traction for its scalability and ease of expressive speech synthesis (Nguyen et al., 2024; Huang et al., 2025; Chen et al., 2025; KimiTeam et al., 2025). SpiritLM (Nguyen et al., 2024), for instance, enhances expressiveness by incorporating style and pitch tokens (Kharitonov et al., 2022; Duquenne et al., 2023), while Step-Audio (Huang et al., 2025) leverages a linguistic tokenizer trained on Paraformer outputs (Gao et al., 2022). The focus on expressivity in these models makes them strong candidates for our visual integration.

3 Motivate the Integration of Visual Cues: Audio-Visual Emotion Recognition

Before diving into Expressive Audio-Visual Language Modeling, we first motivate the study by showing why visual cues, besides its semantic correlation to speech in lip-reading, could be useful. We choose the emotion recognition benchmark, IEMOCAP (Busso et al., 2008), and compare the performance of existing audio-only baselines and our simple audio-visual classification model.

IEMOCAP is made up of videos of two speakers' conversation with strong emotions. We follow prior benchmark EmoBox (Ma et al., 2024b) to train and evaluate models with their released data splits[2]. We follow the findings of EmoBox to use Whisper (Radford et al., 2022) as the audio encoder. For visual information, we directly encode frames with pre-trained Open-MAGVIT2 (Yu et al., 2024; Luo et al., 2025) encoder[3], which has achieved impressive image and video reconstruction performance.

[2] https://github.com/emo-box/EmoBox

[3] https://github.com/TencentARC/SEED-Voken

Model (Modality)	UA (%) ↑	WA (%) ↑	F1 (%)↑
EmoBox (Speech)	73.5	72.9	73.1
SenseVoice-L (Speech)	73.9	75.3	73.2
Ours (Visual)	77.9	79.1	78.7
Ours (Speech+Visual)	**88.2**	**89.2**	**88.5**

Table 1: Comparison of emotion recognition models leveraging different modalities. The performance is evaluated over unweighted average accuracy (UA), weighted average accuracy (WA), and macro F1 score (F1) following SenseVoice (An et al., 2024).

Fusion	Speech	Visual	# Samples
✗	✓	✓	38
✓	✗	✗	119
✓	✓	✗	**646**
✓	✗	✓	**537**

Table 2: Analysis of classification correctness across different models. ✓ indicates correct prediction while ✗ indicates incorrect prediction.

To train our classifier, we encode visual and speech input into features, adapt them through lightweight encoders, and feed them into a Transformer-Encoder model (Vaswani et al., 2017). The features are then pooled and passed into a Feedforward network to predict the emotion label. For details of our model setup, please refer to Appendix B. As shown in Table 1, we compare our classification model with state-of-the-art speech-only models, EmoBox (Ma et al., 2024b) and SenseVoice (An et al., 2024). We find that, by incorporating visual modality, our model achieves superior performance.

We perform an ablation study by grouping samples based on the prediction outcomes of each modality. As shown in Table 2, the fusion model makes more correct predictions via an *ensembling mechanism*. Notably, only a small number of samples are correctly predicted solely by the fusion model (119) or by a single-modality model alone (38). In most cases, the fusion model aligns with the correct single-modality prediction—646 times with speech-only and 537 times with visual-only model—indicating effective modality selection.

These findings indicate that the visual modality provides complementary non-semantic information. Effectively leveraging such visual signals holds promise for enhancing the expressiveness and emotional alignment of speech generation systems.

4 Audio-Visual Language Modeling

To support expressive speech generation, we propose Audio-Visual Language Modeling and explore modality fusion strategies in §4.1. We then discuss our fine-tuning strategy for emotion recognition and expressive speech generation in §4.2.

4.1 AVLM Pre-training

To integrate visual modality into a pre-trained SpeechLM, we leverage self-supervised learning on raw video data (*to distinguish this stage from future fine-tuning experiments, we call this phase AVLM pre-training*). Since the base SpeechLM is already trained on speech and text tokens and employs a Transformer Decoder-only architecture, we aim to inject and align visual information into the existing speech latent space while maintaining next-token prediction as the training objective.

Given speech features $\boldsymbol{s} = (s_1, s_2, \cdots, s_T)$ and visual features $\boldsymbol{v} = (v_1, v_2, \cdots, v_T)$, we propose three audio-visual fusion strategies to obtain audio-visual representations $\boldsymbol{h}$, as visualized in Fig. 2.

DIRECT CONCAT fuses modalities by simply concatenating the time-aligned audio and visual features, which are then projected through a two-layer feedforward network (FFN). The resulting feature is subsequently fed into SpeechLM to generate the speech token sequence $\boldsymbol{s}$.

Both Q-FORMER INFILL and Q-FORMER PREFIX leverage the Q-Former module (Li et al., 2023), which uses a set of query latents to retrieve relevant information from the visual stream. These query latents are initialized as learnable parameters and refined via a cross-attention mechanism that attends to the visual features.

Formally, given a visual stream $\boldsymbol{v}$ and a sequence of query latents $\boldsymbol{q} = (q_1, q_2, \cdots, q_{|\boldsymbol{q}|})$, we first apply sinusoidal positional embeddings to the query latents. They are then passed through a cross-attention module, where each layer contains learnable projection matrices $\boldsymbol{W}_\mathrm{Q}, \boldsymbol{W}_\mathrm{K}, \boldsymbol{W}_\mathrm{V} \in \mathbb{R}^{d\times d}$. Here, d is the shared dimensionality of the query, visual, and speech representation spaces. The query representation is computed using the standard cross-attention formula:

$$\boldsymbol{z} = \mathrm{softmax}\left(\frac{(\boldsymbol{q}\boldsymbol{W}_\mathrm{Q})(\boldsymbol{v}\boldsymbol{W}_\mathrm{K})^T}{\sqrt{d}}\right)\cdot(\boldsymbol{v}\boldsymbol{W}_\mathrm{V}) \quad (1)$$

For Q-FORMER INFILL, we set the number of query latents to match the length of the speech sequence, i.e., $|\boldsymbol{z}| = |\boldsymbol{q}| = |\boldsymbol{s}|$, and randomly replace a portion of the speech representations with query latents. This approach is motivated by AV-HuBERT (Shi et al., 2022), where modality dropout

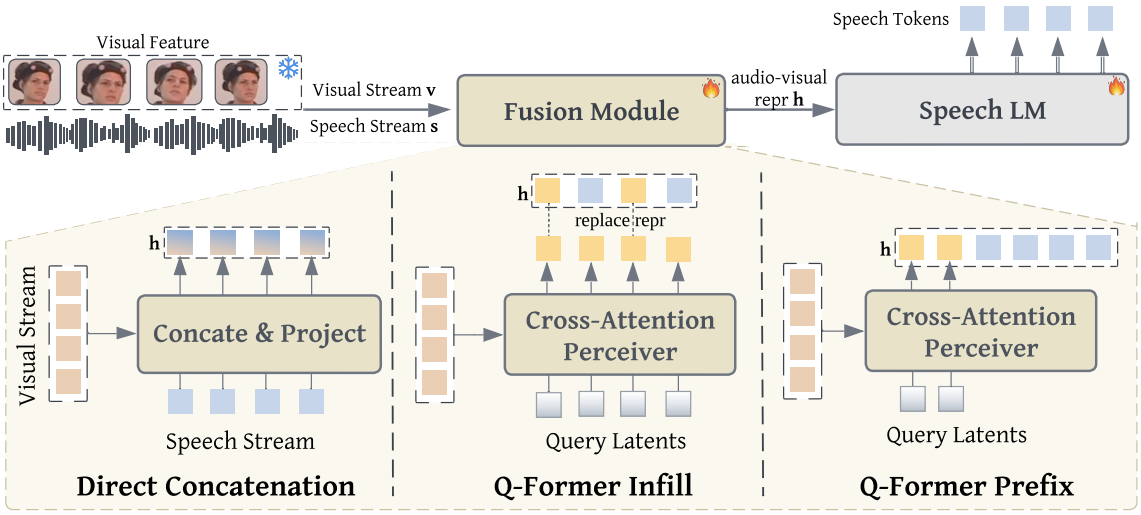

Figure 2: Illustration of the modality fusion strategies explored in the AVLM **pre-training** stage. The fusion module is used to align audio and visual features and feed joint representations to SpeechLM for speech token prediction.

encourages the model to learn robust audio-visual semantic alignment. In Q-FORMER PREFIX, we **dynamically determine the number of query latents** based on the number of visual frames and use them as a prefix prepended to the speech representations.[4] To ensure the model attends to the visual input, we additionally apply attention masking to some speech positions (see §5.1 for details).

To summarize, the audio-visual representation $\boldsymbol{h}$ is obtained with the following equations:

$$\begin{cases} \boldsymbol{h}_i = \text{FFN}\left([\boldsymbol{s}_i \circ \boldsymbol{v}_i]\right) & \text{(CONCAT)} \\ \boldsymbol{h}_i = \boldsymbol{r}_i \cdot \boldsymbol{z}_i + (1 - \boldsymbol{r}_i) \cdot \boldsymbol{s}_i & \text{(INFILL)} \\ \boldsymbol{h}_i = \boldsymbol{z}_i \text{ if } i < |\boldsymbol{z}| \text{ else } \boldsymbol{s}_i & \text{(PREFIX)} \end{cases} \quad (2)$$

where $\boldsymbol{r}_i$ is 1 at position i where we replace the speech with a visual representation and 0 otherwise. $\circ$ denotes concatenation along feature dimension.

Subsequently, our AVLM is pre-trained with the negative log-likelihood (NLL) loss where parameters of the SpeechLM (θ_{lm}) and fusion module (ϕ_{fusion}) are both updated.

$$\mathcal{L}_{\text{pretrain}} = -\sum_{t=1}^{T} \log P(s_t \mid \mathbf{h}_{<t}; \theta_{\text{lm}}, \phi_{\text{fusion}}) \quad (3)$$

In our experiments, we find that Q-FORMER PREFIX achieves the best performance. Therefore, we adopt the Q-FORMER PREFIX AVLM as the base model for fine-tuning in the next section.

[4]In practice, we downsample the visual frames by a factor of 5×.

4.2 AVLM Fine-tuning for Emotion Recognition and Expressive Generation

Given the pre-trained AVLM from §5.1, we fine-tune it on expressive audio-visual conversations annotated with an emotion label. The dataset is derived from IEMOCAP, providing us with an input audio-visual pair, an emotion label, and the output speech. The construction details are shown in the experiment sections and Appendix A, and some example data entries are provided in Table 9.

Instruction-Tuning Prompt

<Text Instruction> ($\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$): Perceive the given visual and audio input. Determine the emotion of the speaker and continue the dialogue with the same emotion.
<Visual Input>: $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N$
<Input Audio>: $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_T$
<Emotion>: $\boldsymbol{e}$ (Happy, Angry, ...)
<Speech Response>: $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M$

As shown in Fig. 3 and colorbox above, we use a multimodal prompt for expressive speech generation. Specifically, we extract visual query representations $\boldsymbol{z} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N\}$ from the input visual features and use them as a prefix to the input speech tokens $\boldsymbol{s} = \{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_T\}$, with $N < T$ due to the compression from Q-FORMER. We then insert the emotion label $\boldsymbol{e}$ to guide generation, followed by the target response speech tokens $\boldsymbol{y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_M\}$.

To fine-tune the model, we minimize the negative log-likelihood (NLL) of the target response tokens $\boldsymbol{y} = \{y_1, \ldots, y_M\}$, conditioned on the multimodal prompt comprising the instruction $\boldsymbol{x}$, visual prefix

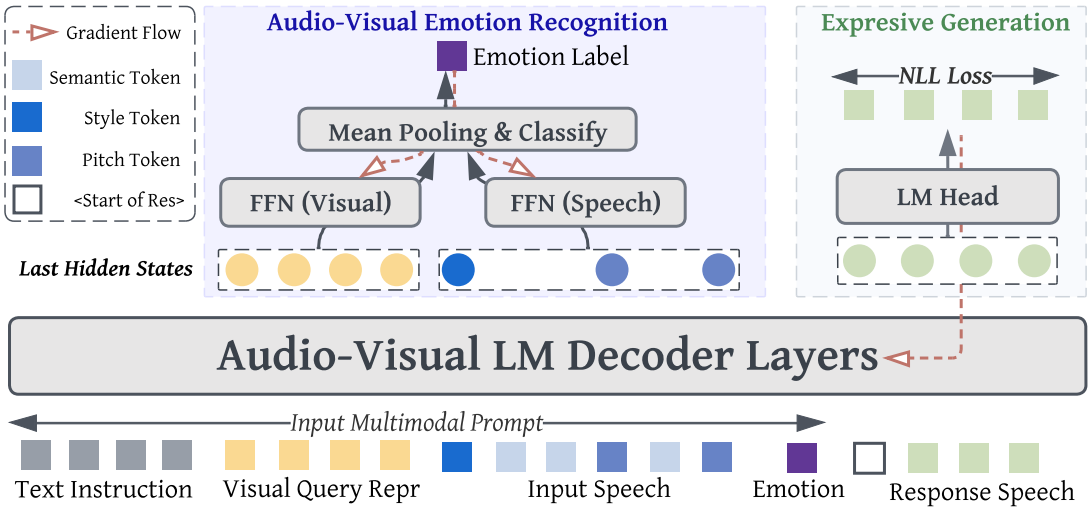

Figure 3: AVLM Multi-task **Fine-tuning** Objectives. The AVLM receives a multimodal input prompt and generates a speech response. An auxiliary emotion recognition module is trained based on transformed visual & speech features.

$\boldsymbol{z}$, input speech $\boldsymbol{s}$, and emotion label $\boldsymbol{e}$:

$$\mathcal{L}_{\text{finetune}} = -\sum_{t=1}^{M} \log P\left(y_t \mid [\boldsymbol{x} \parallel \boldsymbol{z} \parallel \boldsymbol{s} \parallel \boldsymbol{e}]; \theta_{\text{lm}}\right) \tag{4}$$

Here, $\parallel$ denotes the concatenation of input components along the sequence dimension.

Auxiliary Emotion Recognition Task For expressive conversation synthesis using IEMOCAP, we ensure that both input and response speech share the same emotional tone. This is possible as IEMOCAP uses scripted dialogue where speakers of each conversation share similar emotions.

Since autoregressive emotion prediction is unreliable—due to limited data and hallucination issues of decoder-only models—we introduce an auxiliary emotion classification objective. As illustrated in Fig. 3, we extract hidden states corresponding to (1) visual query representations and (2) style/pitch tokens from the SpiritLM-Expressive Tokenizer (details in §5.3). These states are passed through two feedforward networks, concatenated, pooled, and classified using a cross-entropy loss. A stop-gradient is applied to prevent interference with AVLM training. During training, the true emotion label is used while at inference, the emotion label is predicted by the classifier.

5 Experiments

We begin with AVLM pre-training experiments (§5.1), comparing different visual encoders and modality fusion strategies. Next, we fine-tune the AVLM on AVSR tasks (§5.2) to validate the effectiveness of our pre-training. Finally, we fine-tune AVLM for emotion recognition and expressive speech generation (§5.3), achieving superior performance over SpeechLM by utilizing visual cues.

5.1 AVLM Pre-training

Dataset and Preprocessing We use the LRS3 dataset (Afouras et al., 2018), which consists of 433 hours of transcribed audio-visual data, including a 30-hour clean subset. For pre-training, we utilize only the videos from the full 433-hour dataset.

We adopt the data pre-processing pipeline from MuAViC (Anwar et al., 2023), which segments videos into aligned audio-visual chunks of up to 15 seconds based on transcriptions. Speech tokens are extracted using SpiritLM's tokenizer, which produces interleaved tokens of three types: semantic tokens (24.99 fps), style tokens (1 fps), and pitch tokens (12.5 fps). The semantic tokens are discretized HuBERT representations (Hsu et al., 2021), the pitch tokens are derived from a VQ-VAE model trained on the F0 of speech (Polyak et al., 2021), and the style tokens are based on *speechprop* features (Duquenne et al., 2023). For further details, we refer readers to Nguyen et al. (2024).

Model Implementation Details We compare three visual encoders: (1) Open-MAGVIT2 (Yu et al., 2024), (2) VGG-Face2 (Cao et al., 2018), and (3) SMIRK (Retsinas et al., 2024). Open-MAGVIT2 is a capable but expensive encoder as it is trained for image reconstruction. VGG-Face2 is more lightweight and focuses on facial regions through tasks like face recognition. SMIRK is a lightweight neural feature learned through 3D re-

Fusion Module	Visual Feature	PPL ↓
Baseline		
SPEECH-ONLY	N/A	5.6
Fusion: Direct Concatenation		
DIRECT CONCAT	SMIRK	132.0
Fusion: Qformer Variants		
Q-FORMER INFILL	SMIRK	5.8
Q-FORMER PREFIX	SMIRK	**5.5**
Q-FORMER PREFIX	VGG-Face2	8.2
Q-FORMER PREFIX	MAGVIT2	8.6

Table 3: Perplexity (PPL) Comparison across Fusion Modules and Visual Features.

construction based on FLAME (Li et al., 2017). We extract visual features using different encoders, adapt them with lightweight modules, and integrate them into the Q-FORMER for AVLM pre-training (visual encoder comparison details in Appendix C).

For all three fusion modules, we use the same base model, SpiritLM, and apply LoRA fine-tuning (Hu et al., 2021) to the query, key, value, and output projection matrices, with $r = 16$, $\alpha = 32$, and a dropout rate of 0.05. For Q-FORMER INFILL, we replace 50% of the speech tokens with visual query representations. For Q-FORMER PREFIX, we drop attention over a certain percentage of speech tokens (from 0% to 70% as shown in Table 4) to see if such masking helps AVLM become more robust. The details and hyperparameters of our three fusion architectures are provided in Appendix D.

Results We evaluate pre-training performance on the LRS3 test set using perplexity (PPL). First, fixing the fusion method to Q-FORMER PREFIX, we ablate visual encoders. As shown in Table 3, SMIRK achieves the lowest PPL, likely due to its disentangled expressive and jaw features being easier for the model to pick up. **We therefore adopt SMIRK as our default visual encoder for future fine-tuning experiments.**

Next, using SMIRK, we compare fusion strategies. Q-FORMER PREFIX yields the lowest PPL, outperforming both SPEECH-ONLY and other fusion methods. In contrast, DIRECT CONCAT performs worst, suggesting that the direct concatenation and projection of audio-visual features complicate the model's adaptation to the representations.

Among Q-Former variants, Q-FORMER PREFIX consistently performs best. We further evaluate robustness by applying attention masking at inference (10–70%) and training additional Q-FORMER PREFIX variants with various attention masking ratios (Table 4). A 30% speech token dropout from attention computation during training yields the best overall results. Thus, we use the **Q-FORMER PREFIX model trained with 30% masking** as pre-trained base model for all fine-tuning tasks, including AVSR and expressive generation.

Fusion Mode	Train Mask	PPL (↓) @ Attention Mask Ratio				
		0%	10%	30%	50%	70%
SPEECH-ONLY	–	5.56	6.02	7.69	10.27	13.62
INFILL	50%	5.84	6.03	6.60	7.80	9.10
PREFIX	0%	**5.5**	5.87	7.46	9.81	12.0
	30%	5.65	**5.81**	**6.31**	**7.50**	8.88
	50%	5.68	5.87	6.37	7.56	8.81
	70%	5.88	6.03	6.50	7.69	**8.80**

Table 4: Perplexity under varying speech token mask ratios to compare the robustness of fusion modules. Each row is trained with a different speech token mask ratio.

5.2 AVSR Fine-tuning

To further evaluate our pre-trained AVLM, we fine-tune it on the Audio-Visual Speech Recognition (AVSR) task using the 30-hour clean subset of LRS3 and report Word Error Rate (WER). We compare against two baselines: (1) a SPEECH-ONLY SpiritLM model directly fine-tuned on the same 30-hour subset (using speech prefix and text continuation), and (2) an AVLM model with the same architecture as the pre-trained one but fine-tuned from scratch.

AVSR serves as an auxiliary benchmark to assess the quality of visual integration beyond perplexity (used during pre-training). Due to space constraints, we omit modeling details, which largely mirror those in expressive generation (§4.2) by using text transcriptions rather than response speech as targets. Full implementation details are in Appendix E.

We evaluate all models under both clean and noisy conditions. We introduce noise using two methods: (1) SNR noise injection and (2) attention masking. For SNR noise injection, we follow previous studies (Yeo et al., 2025b) by adding white noise to create test audio with varying Signal-to-Noise Ratios (SNR). For attention masking, we randomly mask a certain percentage of speech tokens during attention computation.

Results As demonstrated in Table 5, our pre-trained AVLM achieves the lowest Word Error Rate (WER) in both clean and noisy environments. This indicates that the visual integration via the Q-FORMER PREFIX is effective, and pre-training on a larger dataset is advantageous.

Model	Hours	Clean	WER (%) @ SNR Noise Injection				WER (%) @ Attention Mask Ratio			
			10 dB	5 dB	2 dB	0 dB	10%	30%	50%	70%
SPEECH-ONLY	30	4.43	10.80	18.51	27.92	36.41	34.97	46.04	67.76	82.90
AVLM	30	4.30	9.06	16.97	25.93	32.59	6.53	21.54	52.14	79.86
AVLM	433	**3.50**	**8.58**	**15.18**	**23.90**	**31.50**	**5.44**	**19.10**	**51.70**	**78.90**

Table 5: WER (%) of different models under varying SNR noise levels and speech token mask ratios.

Model	Modality	Hours	WER (%)
Ours			
SPEECH-ONLY (SPIRITLM)	S,T	30	4.43
AVLM (Q-FORMER PREFIX)	S,T	433	3.50
Prior Work (Lip-only)			
AV-HuBERT (Large)	T	433	4.20
MMS-LLaMa	T	433	0.92
LLaMa-AVSR	T	433	0.95

Table 6: WER (%) on clean speech. "Modality" indicates support for speech (S) or text (T) generation.

We also compare our results with previous top-performing models in Table 6. Although models like MMS-LLaMa (Yeo et al., 2025b) and LLaMa-AVSR (Cappellazzo et al., 2025) report lower WER, the differences primarily stem from our choice of base model and visual features, rather than the modeling approach itself.

Both MMS-LLaMa and LLaMa-AVSR utilize audio encoders like Whisper and focus on visual features from the lip region, limiting their models to the AVSR task. In contrast, our base model, SpiritLM, supports expressive speech generation by encoding speech into units, which inevitably loses information from speech tokenization and result in higher WER. Moreover, our visual features encompass the full face, not just the lip region, to facilitate expressive generation but at the cost of less effective semantic alignment for tasks like AVSR.

5.3 Emotion Recognition and Expressive Speech Generation

Dataset Following the approach in §4.2, we fine-tune our pre-trained AVLM using multimodal prompts consisting of visual input, input and response audio, and an emotion label. The dataset is built from IEMOCAP (Busso et al., 2008), which provides expressive video dialogues and emotion labels. However, many original responses are very short (e.g., acknowledgments like "yeah"). To address this, we use GPT-4 (OpenAI et al., 2024)[5] to rewrite conversations into longer, more detailed responses. We then generate corresponding audio using Step-Audio-TTS-3B (Huang et al., 2025), an expressive TTS model supporting voice cloning and style control. Additional dataset construction details are provided in Appendix A.

Model Implementation We initialize our model from the pre-trained AVLM (§5.1) using the Q-FORMER PREFIX architecture with the SMIRK visual encoder, and fine-tune it with LoRA using the same hyperparameters as pre-training. The emotion classifier is trained separately on hidden states from visual queries and style/pitch tokens posii (§4.2). For the speech-only baseline, we fine-tune a pre-trained SpiritLM model with LoRA, excluding visual inputs. Its emotion classifier uses only the style and pitch token hidden states.

Evaluation For emotion recognition, we evaluate performance by comparing the predicted and ground truth emotion labels, computing both accuracy and F1-score for four emotion categories: Happy, Sad, Angry, and Neutral. For expressive speech generation, we use fine-tuned model to generate speech tokens and synthesize speech through SpiritLM's Tokenizer. Subsequently, we use a third-party model, Qwen2-Audio (Chu et al., 2024), which excels at audio understanding and emotion recognition, *to predict emotion labels from the generated speech* (see Appendix D for details). We then compute accuracy and F1-score for the same four-way classification. For reference, we have also uploaded audio of some generated responses in our ARR supplementary material.

Results As shown in Table 7, our AVLM model outperforms the SPEECH-ONLY model on both emotion recognition and speech generation tasks. Note that, both models achieve better performance on emotion recognition, which is expected since this task directly optimizes for emotion label prediction. In contrast, speech generation is more challenging: the decoder-only model may hallucinate, producing speech responses that fail to convey the intended emotion. Nevertheless, across both tasks, our

[5]We use the "gpt-4o-2024-11-20" snapshot

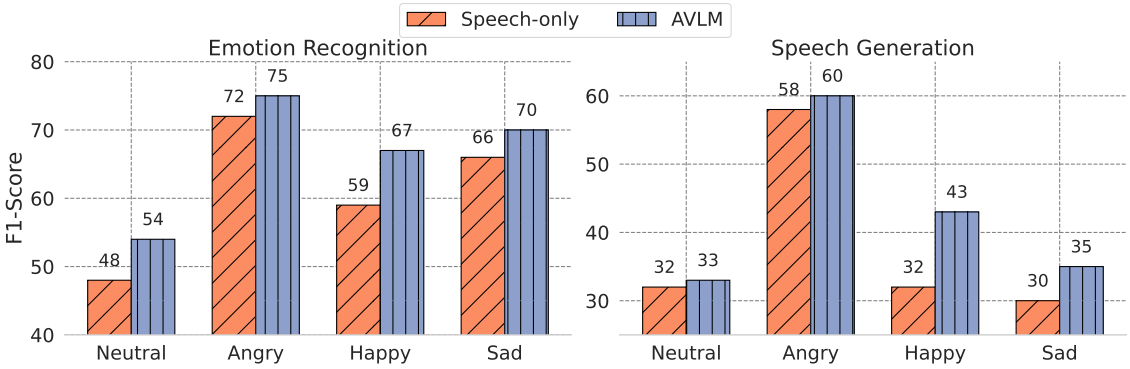

Figure 4: F1-score of each emotion class for AVLM and SPEECH-ONLY model.

Model	UA (%)	WA (%)	F1 (%)
Emotion Recognition			
SPEECH-ONLY	62.8	61.3	61.3
AVLM	**67.2**	**67.1**	**66.2**
Speech Generation			
SPEECH-ONLY	43.79	38.85	38.39
AVLM	**46.03**	**42.99**	**42.49**

Table 7: Performance comparison of AVLM and Speech-only model in terms of unweighted accuracy (UA), weighted accuracy (WA), and macro F1-score (F1).

AVLM model consistently surpasses the SPEECH-ONLY baseline, highlighting the benefit of visual guidance. In Fig. 4, we further break down the F1-scores by emotion category, observing a similar trend where AVLM always outperforms its speech-only counterpart. Among the emotions, Neutral and Sad are generally harder to predict than Happy and Angry, likely because they convey weaker emotional cues. On the contrary, Happy and Angry are associated with stronger facial expressions and higher vocal arousal, making it easier for the model to predict or generate speech of similar properties.

Emotion Controllability Analysis We investigated whether the emotion of generated speech could be controlled by modifying the predicted emotion label in the prompt. For example, if the original predicted emotion is "angry", we manually change it to "happy" and then evaluate the emotion of the newly generated speech. We measure the number of instances where the generated emotion successfully shifts from the original to the altered label and visualize the results in a heatmap. Ideally, we expect our AVLM to demonstrate strong controllability, generating speech that appropriately reflects the emotion specified in the prompt.

As shown in Fig. 5, simply changing the emotion label in the prompt—without using in-context learning—rarely alters the emotional tone of the

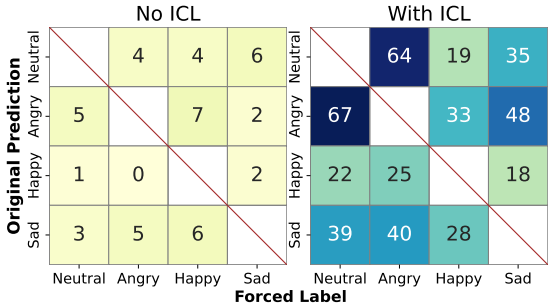

Figure 5: Heatmaps of emotion controllability analysis for AVLM. The number indicates the amount of generated speech that changes from its original emotion to the forced emotion label. ICL: In-context Learning.

generated speech. This suggests that the model primarily relies on input speech and visual cues to guide emotion, rather than the emotion label itself. This behavior is expected, given the limited training data and the fact that input and response audio in our dataset are designed to share the same emotion.

To enhance controllability, we introduce in-context learning by including one demonstration per emotion in the prompt (see example in Appendix D). With this addition, AVLM becomes more responsive to changes in the emotion label, as shown in the right panel of Fig. 5.

These findings indicate there is still large room for improving the emotion controllability, as the model after fine-tuning is not able to easily alter the speech style by conditioning on different emotion labels. We believe that enhancing data quality by creating larger and more diverse expressive audio-visual dialogue is crucial to addressing this issue, and we plan to explore this in future work.

6 Conclusion

We present a framework for Audio-Visual Language Modeling that effectively incorporates visual cues into SpeechLM. Following empirical exploration

of visual encoders and modeling architectures, we employ a prefix-based Query-Former with SMIRK visual features for AVLM. Our system, through leveraging visual modality, outperforms speech-only baselines in both Audio-Visual Speech Recognition and Expressive Speech Generation.

Acknowledgement

We express our profound appreciation to anonymous reviewers for their helpful suggestions. We also thank Yue Xu, Tianjian Li,Jingyu Zhang for their valuable suggestions to enhance the presentation of this work.

Limitations and Broader Impact

One limitation of our work is the limited availability of expressive audio-visual dialogue data for fine-tuning the pre-trained AVLM. Although we synthesize part of our dataset, the construction pipeline is based on IEMOCAP, which contains only a few hours of recordings. Collecting more diverse and expressive dialogue data could further enhance the controllability and quality of the generated speech.

Another limitation is that our evaluation primarily focuses on the emotional expressiveness of the generated speech, reflecting our goal of modeling emotion-aware speech generation from audio-visual inputs. While we demonstrate that visual cues can significantly improve emotional alignment, other aspects of speech quality—such as helpfulness, factual accuracy, and coherence—are not evaluated in this study. Finally, the decoding process in our Q-FORMER PREFIX architecture requires processing the visual stream before the speech stream, which can introduce latency and leave room for further efficiency improvements.

Ethical Consideration: While our work advances the generation of emotionally expressive and natural-sounding speech, it also raises potential risks of misuse. Specifically, models capable of synthesizing human-like emotional speech may be exploited for deceptive purposes, such as impersonation, social engineering scams, or the spread of misinformation. To mitigate such concerns, we encourage future work to incorporate safeguards such as synthetic speech watermarking, controlled access to fine-tuned models, and explicit labeling of machine-generated content. Responsible deployment and continuous risk assessment are essential as these technologies mature.

Our experiments are conducted on publicly available datasets under appropriate licenses. The LRS3 dataset is released under the Creative Commons Attribution 4.0 International License and is available for research purposes. The IEMOCAP dataset is also released for academic research use under its license agreement. We use both datasets in accordance with their respective terms to ensure compliance with data usage policies. The synthesized artifacts for fine-tuning are not currently distributed to avoid potential misuse and to respect the licensing constraints from IEMOCAP (Busso et al., 2008). We are planning to release the dataset if permission is granted from the IEMOCAP data provider. We believe responsible handling of such synthetic content is critical to ensuring ethical standards in expressive speech research.

We also acknowledge the use of AI assistants (e.g., GitHub Copilot, ChatGPT) to support the development of our research. These tools were used to assist with code implementation, debugging, documentation, and drafting or refining written content. All substantive research contributions, model design decisions, and experimental results were generated and verified by the authors. We ensured that the use of AI tools complied with academic integrity standards and that any generated content was critically reviewed and edited to maintain originality and correctness.

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Lrs3-ted: a large-scale dataset for visual speech recognition. *Preprint*, arXiv:1809.00496.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, and 14 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *Preprint*, arXiv:2407.04051.

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *Preprint*, arXiv:2303.00628.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2017. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

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

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. *Preprint*, arXiv:1710.08092.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. 2025. Large language models are strong audio-visual speech recognition learners. *Preprint*, arXiv:2409.12319.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, Yabin Li, Xiang Lv, Jiaqing Liu, Haoneng Luo, Bin Ma, Chongjia Ni, Xian Shi, Jialong Tang, Hui Wang, and 17 others. 2025. Minmo: A multimodal large language model for seamless voice interaction. *Preprint*, arXiv:2501.06282.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, pages 1086–1090.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *Preprint*, arXiv:2412.10117.

Paul-Ambroise Duquenne, Kevin Heffernan, Alexandre Mourachko, Benoît Sagot, and Holger Schwenk. 2023. SONAR EXPRESSIVE: Zero-shot Expressive Speech-to-Speech Translation. Working paper or preprint.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *Preprint*, arXiv:2210.13438.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2063–2067. ISCA.

Alexandros Haliassos, Rodrigo Mira, Honglie Chen, Zoe Landgraf, Stavros Petridis, and Maja Pantic. 2024. Unified speech recognition: A single model for auditory, visual, and audiovisual inputs. In *Advances in Neural Information Processing Systems*, volume 37, pages 139673–139699. Curran Associates, Inc.

HyoJung Han, Mohamed Anwar, Juan Pino, Wei-Ning Hsu, Marine Carpuat, Bowen Shi, and Changhan Wang. 2024. XLAVS-R: Cross-lingual audio-visual speech representation learning for noise-robust speech perception. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12896–12911, Bangkok, Thailand. Association for Computational Linguistics.

Joanna Hong, Minsu Kim, Jeong Yun Choi, and Yong Man Ro. 2023. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18783–18794.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, Peng Liu, Ruihang Miao, Wang You, Xi Chen, Xuerui Yang, Yechang Huang, Yuxiang Zhang, Zheng Gong, Zixin Zhang, and 126 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *Preprint*, arXiv:2502.11946.

Teerapat Jenrungrot, Michael Chinen, W. Bastiaan Kleijn, Jan Skoglund, Zalán Borsos, Neil Zeghidour, and Marco Tagliasacchi. 2023. Lmcodec: A low bitrate speech codec with causal transformer models. *Preprint*, arXiv:2303.12984.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6–7):907–929.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. 2025. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *Preprint*, arXiv:2409.04410.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2024a. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26:776–788.

Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. 2024b. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. *ArXiv*, abs/2406.07162.

Albert Mehrabian and James A Russell. 1974. *An Approach to Environmental Psychology*. MIT Press.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. Spirit-lm: Interleaved spoken and written language model. *Preprint*, arXiv:2402.05755.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *Preprint*, arXiv:2104.00355.

R. Gnana Praveen, Patrick Cardinal, and Eric Granger. 2023. Audio–visual fusion for emotion recognition in the valence–arousal space using joint cross-attention. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(3):360–373.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. *Preprint*, arXiv:1506.02640.

George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 2024. 3d facial expressions through analysis-by-neural-synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fabien Ringeval, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE.

Andrey V. Savchenko. 2022. Hsemotion: High-speed emotion recognition library. *Software Impacts*, 14:100433.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *Preprint*, arXiv:2201.02184.

Weiting Tan, Hirofumi Inaguma, Ning Dong, Paden Tomasello, and Xutai Ma. 2024. Ssr: Alignment-aware modality connector for speech language models. *Preprint*, arXiv:2410.00168.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024a. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

Haobin Tang, Xulong Zhang, Ning Cheng, Jing Xiao, and Jianzong Wang. 2024b. Ed-tts: Multi-scale emotion modeling using cross-domain emotion diarization for emotional speech synthesis. *Preprint*, arXiv:2401.08166.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shijun Wang, Jón Guðnason, and Damian Borth. 2023. Learning emotional representations from imbalanced speech data for speech emotion recognition and emotional text-to-speech. *Preprint*, arXiv:2306.05709.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and Yu Wu. 2023. On decoder-only architecture for speech-to-text and large language model integration. *Preprint*, arXiv:2307.03917.

Jeong Hun Yeo, Minsu Kim, Chae Won Kim, Stavros Petridis, and Yong Man Ro. 2025a. Zero-avsr: Zero-shot audio-visual speech recognition with llms by learning language-agnostic speech representations. *Preprint*, arXiv:2503.06273.

Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, and Yong Man Ro. 2025b. Mms-llama: Efficient llm-based audio-visual speech recognition with minimal multimodal speech tokens. *Preprint*, arXiv:2503.11315.

Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. 2024. Language model beats diffusion – tokenizer is key to visual generation. *Preprint*, arXiv:2310.05737.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Connecting speech encoder and large language model for asr. *Preprint*, arXiv:2309.13963.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *Preprint*, arXiv:2305.11000.

Supplementary Material

Appendix Sections	Contents
Appendix A	Expressive Conversation Dataset Construction from IEMOCAP
Appendix B	IEMOCAP Emotion Recognition Experiment Setup
Appendix C	Comparison of Visual Encoders for AVLM Pre-training
Appendix D	Audio-Visual Language Modeling Implementation Details
Appendix E	Audio-Visual Speech Recognition Modeling Details

A Expressive Audio-Visual Dataset Construction

To showcase the benefit of our visual integration, we synthesize data from IEMOCAP (Busso et al., 2008) to create an expressive data collection that contains video input, response audio, and emotion label. Note that although the original IEMOCAP dataset already contains data format that satisfies this requirement, we find the conversation are not good enough to fine-tune our AVLM because a lot of the dialogue turns are too short to be useful. Some example conversation from original dataset are shown below in Table 8.

Example	Speaker A	Speaker B
1	Do you have your forms?	Yeah.
2	Who told you to get in this line?	You did.
3	I'm a big fan of the Wheel of Fortune quarters. But it just costs so much. I don't know.	Oh.
4	Well so now we're married. We have cat children.	Awesome.

Table 8: Original short conversations motivating synthetic augmentation

In order to fine-tune our AVLM for better response generation, we perform conversatoin re-writing to generate more detailed response using GPT-4 (OpenAI et al., 2024). The prompt we use is following:

Conversation Rewriting Prompt

```
<Instruction>: Your task is to continue the following conversation naturally, taking into account
both the conversation history and the provided emotional tone. You will generate a response
from the perspective of a [speaker_gender] speaker, ensuring that your tone reflects a [emotion]
emotion. An example response is also provided for reference; if it is very short, please expand
it to one or two sentences.

<Conversation History>: [Previous Conversation]
<Current Speaker's Emotion>: [Emotion Label]
<Example Response>: [Original Response Text]

Now, generate a natural and concise response that continues the dialogue using all of the above
information. Keep your answer to approximately two sentences. Output the response only.
```

The rewritten conversation will have better length control as we prompt the model to keep the response to approximately two sentences. Since the response also takes into account all conversation history from previous turns and the current speaker gender and emotion, it will greatly improve the quality of the dialogue for training our AVLM. Then we need to synthesize audio from the re-written conversations and we choose Step-Audio-TTS-3B (Huang et al., 2025) to because it supports voice cloning and expressive speech synthesis with natural language style control. We feed the autoregressive TTS model with the following prompt: `[Emotion] [Text Response]` along with a speaker voice clip (from the current conversation) for the model to clone.

After the synthesis process above, we get 4859 conversation pairs and we split it by 80:10:10 for train:dev:test. During our construction, we use five emotion labels: Happy, Sad, Angry, Neutral, and Frustrated, as these are the most common emotions from the original dataset, and we filter out all

conversation with other emotion labels. However, we also realize that the emotion Angry and Frustrated are very similar so we merge them into one class to only have Angry emotion. Lastly, we process the visual stream by first cropping the face region with YOLO (Redmon et al., 2016)[6] and then extract its SMIRK feature. We process the audio stream by SpiritLM's expressive tokenizer to produce semantic, style, and pitch tokens. An example of crafted data is shown in Table 9.

Field	Content
Example A	
Emotion	Sad
Input Visual	SMIRK features extracted at 25 fps
Input Text	No, I know but—I know I don't make you happy.
Response Text	It's just that sometimes it feels like we're stuck in this place, and I don't know how to find my way back to feeling good again. I wish things were different, but I don't want to push you away either.
Input Units	[St48][Pi40][Hu109][Hu118][Hu80][Pi44][Hu42][Hu112][Hu342][Hu490][Hu356][St48][Hu136][Pi56][Hu1][Hu94][Pi44][Hu114]...
Response Units	[St28][Pi9][Hu140][Pi37][Hu482][Hu465][Hu178][Pi21][Hu7][Hu132][Pi49][Hu397][Pi8][Hu145][Hu248][Hu493][Hu383][Hu30]...
Example B	
Emotion	Angry
Input Visual	SMIRK features extracted at 25 fps
Input Text	Do you want to – do you want to leave us behind? I mean, I don't understand why you have to do this.
Response Text	I wish it were different, but this is something I committed to long before any of this happened. It's not about wanting to leave you behind; it's about fulfilling my responsibility.
Input Units	[St71][Pi21][Hu267][Pi52][Hu7][Pi0][Hu469][Hu118][Pi47][Hu410][Hu333][Hu156][Pi21][Hu139][Pi9][Hu359][Pi5][Hu343]...
Response Units	[St71][Pi41][Hu255][Hu437][Hu360][Hu35][Pi46][Hu169][Pi25][Hu70][Hu431][Pi52][Hu389][Hu249][Pi54][Hu95]...
Example C	
Emotion	Happy
Input Visual	SMIRK features extracted at 25 fps
Input Text	Great. That makes us all happy here at D.S.L. Extreme. Is there anything else I can help you with?
Response Text	No, I think that's everything for now! You've really made my day brighter. I'm excited to stick with D.S.L. Extreme!
Input Units	[St22][Pi59][Hu7][Pi20][Hu171][Pi59][Hu118][Hu161][Pi48][Hu71][Hu142][Pi46][Hu251][Hu175][Pi41][Hu7][Hu314][Pi40]...
Response Units	[St69][Pi2][Hu278][Hu80][Pi54][Hu112][Hu81][Pi52][Hu159][Pi11][Pi30][Hu289][Hu126][Pi2][Hu362][Pi30][Hu56][Pi2][Hu7]...
Example D	
Emotion	Neutral
Input Visual	SMIRK features extracted at 25 fps
Input Text	She's actually from England, I think. Well she- she spoke English but she had an accent. I didn't even ask.
Response Text	I guess that works out for both of you then. Have you thought about what it would be like to actually go through with the marriage?
Input Units	[St28][Pi57][Hu475][Pi20][Pi35][Pi47][Hu28][Pi54][Pi52][Pi47][Hu49][Pi35][Hu391][Hu245][Pi13][Hu227][Hu355]...
Response Units	[St71][Pi20][Hu278][Hu35][Pi35][Hu130][Pi30][Hu169][Pi56][Hu459][Hu131][Pi47][Hu197][Hu128][Pi5][Hu424][Pi11][Hu368]...

Table 9: Examples of processed data entries of different emotions. The input text and response text are only shown for reference but not used during AVLM fine-tuning experiments.

The synthesized artifacts for fine-tuning are not currently distributed to avoid potential misuse and to respect the licensing constraints from IEMOCAP (Busso et al., 2008). We are planning to release the dataset if permission is granted from the IEMOCAP data provider. We believe responsible handling of such synthetic content is critical to ensuring ethical standards in expressive speech research.

[6] We use publicly available ckpt: yolov8l-face-lindevs.pt

B IEMOCAP Emotion Recognition Experiment Setup

In §3, we demonstrate the benefit of incorporating visual modality for emotion recognition. Below, we detail our experimental setup:

We follow the data split protocol from EmoBox (Ma et al., 2024b), using the publicly available splits at https://github.com/emo-box/EmoBox/tree/main/data/iemocap. Visual frames are encoded using the Open-MAGVIT2 encoder, while speech signals are processed with *Whisper-Large-V3*. A lightweight adaptation module is then applied to map both speech and visual features into 512-dimensional representations for each frame.

To adapt the visual features from Open-MAGVIT2, which produce tensors of shape $(Z = 18, H = 32, W = 32)$, we stack three residual blocks, each consisting of a normalization layer followed by a 2D convolutional neural network (CNN). The resulting output is flattened and projected into a 512-dimensional feature vector. Speech features are adapted in a similar manner, but using 1D-CNNs instead of 2D-CNNs.

The adapted speech and visual features are then concatenated and fed into a Transformer encoder consisting of 6 layers with a hidden size of 512. We experimented with concatenating along both the temporal and feature dimensions but observed no significant difference in performance. Therefore, we adopt feature-dimension concatenation to keep the sequence length shorter. The resulting features are mean-pooled and passed through a two-layer feedforward network to predict the emotion label. We train the model until convergence, using a learning rate of 3×10^{-5}, a batch size of 16, and gradient accumulation over 4 steps.

To evaluate classifier performance, we follow prior work (Ma et al., 2024b; An et al., 2024) and report the unweighted average accuracy (UA), weighted average accuracy (WA), and macro F1 score, as shown in Table 1. The UA, WA, and F1 metrics are computed using the *accuracy_score* and *balanced_accuracy_score* functions from the *sklearn.metrics* package.

C Visual Encoder Comparison

We initially selected MAGVIT2 (using the public checkpoint from Open-MAGVIT2) for its rich visual representations. However, MAGVIT2 encodes each 256×256 frame into a high-dimensional feature of size $18 \times 32 \times 32 = 18{,}432$, with 18 channels and spatial dimensions downsampled by a factor of 8. This high dimensionality introduces significant computational overhead and poses learning challenges, particularly given the limited size of our dataset (fewer than 500 hours from LRS3). To mitigate these issues, we explored more lightweight visual encoders: VGG-Face2 and SMIRK.

VGG-Face2, trained for face recognition, encodes each frame into a 512-dimensional vector. SMIRK, trained for 3D face reconstruction using FLAME (Li et al., 2017), disentangles facial attributes into components such as shape, expression, jaw, eyelid, pose, and camera parameters. Among these, we use only the expressive (55-dimensional) and jaw (3-dimensional) features, which are most relevant for capturing emotion and lip movements.

Since the outputs of these visual encoders differ in shape, we design specialized adapters before feeding the features into the Q-FORMER. For MAGVIT2, we follow the setup in Appendix B, stacking 2D convolutional layers with residual connections to reduce the spatial dimensions and project the encoding into a 512-dimensional feature.

For VGG-Face2, whose outputs are already 512-dimensional, we apply a simple two-layer feedforward network that first expands the feature to 1024 dimensions before projecting it back to 512 dimensions. For SMIRK, we linearly transform the expressive parameters into a 128-dimensional vector and the jaw parameters into a 32-dimensional vector. These are concatenated and passed through a two-layer feedforward network to produce a 256-dimensional visual representation.

Since SMIRK features can sometimes be distorted, we leverage its pose parameters to filter out video clips with excessive horizontal head rotation. Let the predicted pose vector be $\boldsymbol{r}_{\text{pose}} = (r_x, r_y, r_z)$, initially represented as a rotation vector. We convert it into a rotation matrix $\boldsymbol{R}$ using Rodrigues' rotation formula:

$$\boldsymbol{R} = \mathbf{I} + \sin(\theta)\mathbf{K} + (1 - \cos(\theta))\mathbf{K}^2 \tag{5}$$

where $\theta = |\boldsymbol{r}_{\text{pose}}|$ and $\mathbf{K}$ is the skew-symmetric matrix of the unit vector $\mathbf{u} = \frac{\boldsymbol{r}_{\text{pose}}}{\theta}$, with $\mathbf{I}$ denoting the identity matrix. The resulting matrix $\boldsymbol{R}$ is then converted to Euler angles (ϕ, θ, ψ), corresponding to yaw, pitch, and roll. We use the yaw angle

$$\phi = \text{arctan2}(-R_{31}, R_{33})$$

(where R_{ij} are the elements of the rotation matrix $\boldsymbol{R}$) to quantify horizontal head rotation and filter out clips whose average yaw angle exceed 30%, resulting in about 10% drop in the training data.

As shown in Table 3, we compare the three visual encoders using the Q-FORMER PREFIX architecture and observe that SMIRK outperforms both MAGVIT2 and VGG-Face2. We attribute this superior performance to SMIRK's ability to produce disentangled and emotionally relevant features, thereby facilitating more effective learning.

D Audio-Visual Language Modeling Implementation Details

Training We explored three modality fusion strategies. In the DIRECT CONCAT approach, the encoded visual features are first projected to 256 dimensions and then concatenated with the 4096-dimensional speech features from SpiritLM, resulting in a combined 4352-dimensional representation. This concatenated vector is subsequently projected back to 4096 dimensions through a two-layer feedforward network.

For the Q-FORMER PREFIX and Q-FORMER INFILL methods, we leverage query latents to retrieve relevant visual information using a 6-layer cross-attention module. Before computing cross-attention, we add positional embeddings to the query latents to encode temporal information. In Q-FORMER PREFIX, we apply a compression ratio of 5 to downsample the visual input (i.e., for a one-second video at 25 frames per second, we generate 5 query representations to serve as the prefix). In contrast, in Q-FORMER INFILL, the query latents are sized to match the number of speech tokens; after retrieval from the visual stream, 50% of the speech representations are replaced with the retrieved visual queries.

We adopt LoRA for training our AVLM, using hyperparameters $r = 16$, $\alpha = 32$, and a dropout rate of 0.05. The LoRA parameters are optimized with a learning rate of 3×10^{-5} and 1000 warm-up steps. For training the visual adapter (described in Appendix C), we use a larger learning rate of 1×10^{-4}. The model is trained with a batch size of 8, applying gradient accumulation every 2 steps.

For AVLM fine-tuning on speech generation, we adjust the batch size to 5 due to memory constraints. For the emotion classifier, we train it with learning rate $3e^{-4}$ and for the LoRA parameters, we use learning rate $5e^{-5}$ with no warmup steps. Throughout all experiments, we use $4\times$ A100 80G GPUs and apply Adam Optimizer with $\beta_1 = 0.99, \beta_2 = 0.999$.

Inference During inference, we prepare a multimodal prompt using the text instruction, visual input, and audio input as shown in the zero-shot prompt below:

Zero-Shot Prompt

<Text Instruction>: Perceive the given visual and audio input. Determine the emotion of the speaker and continue the dialogue with the same emotion.
<Visual Input>: $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N$
<Input Audio>: $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_T$

Then, we infer the emotion from our trained audio-visual emotion classifier module, and insert the emotion into the prompt. Lastly, the AVLM model generates speech continuation based on this multimodal input prompt.

For In-context Learning (ICL) analysis, we use a different prompt that contains an example of each emotion, as shown in the prompt below. As discussed in §5.3, with prompt demonstrations, the generated speech better aligns with the emotion label compared to zero-shot prompting, indicating that the model can learn from context to adapt its emotion during expressive generation. This also suggests room for further improvements, as zero-shot setting is not able to achieve good emotion-controllability itself.

In-context Learning Prompt

<Text Instruction>: Perceive the given visual and audio input. Determine the emotion of the speaker and continue the dialogue with the same emotion. Below are some examples:

<Audio-Visual Conversation Input>: visual queries;input speech units
<Emotion>: happy
<Response>: response speech units

<Audio-Visual Conversation Input>: visual queries;input speech units
<Emotion>: sad
<Response>: response speech units

<Audio-Visual Conversation Input>: visual queries;input speech units
<Emotion>: angry
<Response>: response speech units

<Audio-Visual Conversation Input>: visual queries;input speech units
<Emotion>: neutral
<Response>: response speech units

Now continue the following Audio-Visual conversation input with specified emotion
<Visual Input>: $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N$
<Input Audio>: $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_T$
<Emotion>: {emotion label}
<Response>:

To control the structure of generated speech token sequences, we employ a custom decoding strategy using nucleus sampling with temperature, combined with a logits processor. Specifically, we set the generation temperature to 0.8 and use top-p sampling with $p = 0.95$, generating up to 300 new tokens with sampling enabled. Additionally, we apply a LogitProcessor to enforce structural constraints on the token types during generation.

The processor masks out all non-speech tokens and selectively restricts transitions between different token types. Speech tokens are grouped into three disjoint ranges: style tokens, pitch tokens, and semantic tokens. At the first decoding step, only style tokens are allowed. Thereafter, we prohibit consecutive style or pitch tokens by dynamically masking their respective ranges based on the previously generated token type. No such constraint is applied to semantic tokens, allowing them to appear in sequence. This decoding logic helps the model maintain a well-formed token sequence that can be converted back to expressive speech waveforms through pre-trained speech decoders.

To detect the emotion from the generated speech, we perform evaluation with Qwen2-Audio. We prompt it with following text: *What's the emotion of the audio? Only output the emotion label from the following list: Happy, Sad, Angry, Frustrated, Neutral* and provide the synthesized audio. The predicted labels are then used to compute accuracy and F1-score against the ground-truth emotion labels.

E AVSR Finetuning Details

In this section, we elaborate on the modeling and experimental setup used for the AVSR (Audio-Visual Speech Recognition) task, which was omitted from the main text due to space limitations. Our fine-tuning strategy mirrors the multimodal prompting approach described in §5.3, but with one key difference: instead of generating a response in speech, the model is trained to predict transcribed text tokens, as illustrated in Fig. 6. The instruction-tuning prompt used for AVSR adapts accordingly to reflect this new objective. For the speech-only baseline, we directly fine-tune SpiritLM with the same multimodal prompt, except that visual representations are not provided.

Instruction-Tuning Prompt

<Text Instruction>: Transcribe the following audio visual clip:
<Visual Input>: $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N$
<Audio Input>: $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_T$
<Transcription>: (some text tokens)

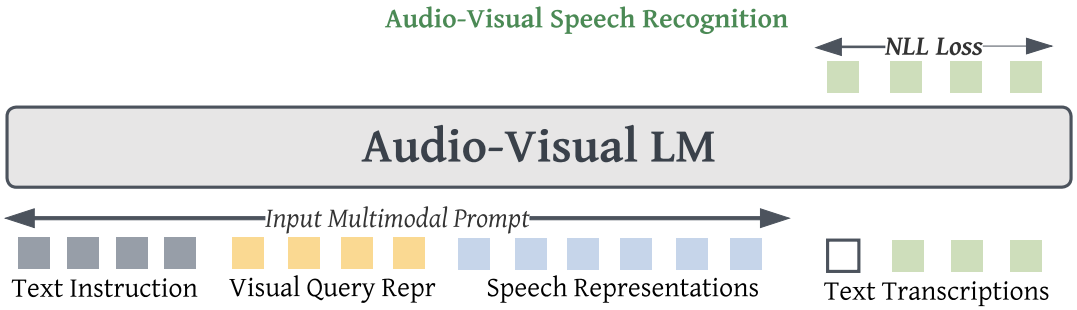

Figure 6: Audio-Visual Speech Recognition Fine-tuning based on pre-trained AVLM model.

We fine-tune our AVLM on the LRS3 dataset using LoRA-based parameter-efficient updates. The training configuration aligns closely with that used for expressive speech generation. Specifically, for the baseline AVLM variant without pretraining, we apply a learning rate of 1×10^{-4} to the fusion modules and 3×10^{-5} to the LoRA parameters within SpeechLM. In contrast, for the pretrained AVLM, we freeze the fusion modules and update only the LoRA parameters. Fine-tuning is conducted with a batch size of 24, using gradient accumulation over 8 steps and 1,000 warmup steps.

To encourage the model to leverage visual input, we adopt a speech masking strategy during both training and evaluation. For a specified dropout ratio, we randomly sample a subset of positions from the input speech token sequence and prevent attention from attending to those positions. This forces the model to rely more heavily on visual cues when performing recognition (results shown in Table 5).