Open Your Model's Eyes: Video and Context-Aware Multimodal Backchannel Prediction

Min-Jae Kim[1,*], **Jun-Yeong Moon**[1,*], **Mujeen Sung**[2,†], **Gyeong-Moon Park**[1,†]
[1]Korea University, Seoul, Republic of Korea
[2]Kyung Hee University, Yongin, Republic of Korea
{minjaekim, moonjunyyy, gm-park}@korea.ac.kr, mujeensung@khu.ac.kr

Abstract

Backchannels, which signal listener states like empathy and understanding, are fundamental to natural human interaction. However, current approaches rely solely on audio and text. This omits crucial visual cues, such as facial expressions and gestures, as well as broader conversational contexts, which are necessary for accurate prediction. In this paper, we introduce Context-Aware Multimodal Alignment for Backchannel Prediction (CAMA-BC), a novel framework that leverages visual information through Multi-Layer Multimodal Alignment (MMA). Our alignment process comprises two stages. First, Context Alignment (MMA-CA) utilizes unlabeled dialogues with videos to capture conversational contexts. Next, Backchannel Alignment (MMA-BA) fine-tunes the representations specifically for backchannel prediction. Experimental results show that CAMA-BC significantly outperforms both existing methods and simple multimodal baselines, with particular effectiveness in recognizing complex backchannels such as empathy.

1 Introduction

Human conversation is a dynamic interplay involving continuous signal exchange to co-construct meaning and maintain mutual understanding. Central to this interaction is **backchanneling**: the production of brief yet essential conversational responses that provide real-time feedback on comprehension, engagement, and affective states of interlocutors (Yngve, 1970). From the initial effort that used simple linear classification (Kawahara et al., 2016), backchannel prediction has emerged as a critical component in natural language processing (NLP) (Ortega et al., 2020; Jang et al., 2021; Ortega et al., 2023).

Existing backchannel prediction methods have relied exclusively on linguistic information, such as textual and acoustic signals. Those approaches systematically overlook the visual modality, which is essential for improving comprehension and facilitating mutual understanding in human interactions (Krauss et al., 1996; Scherer, 2013; Mandal, 2014) because of its richness in facial expressions, gestures, and pre-utterance movements. Our preliminary investigation in Table 1 reveals that incorporating video features from VideoMAE (Tong et al., 2022) through simple concatenation yields only modest improvements. While this confirms that visual information contains valuable signals for backchannel prediction, it simultaneously reveals a critical paradox: naive integration approaches fail to unlock the potential of visual modality.

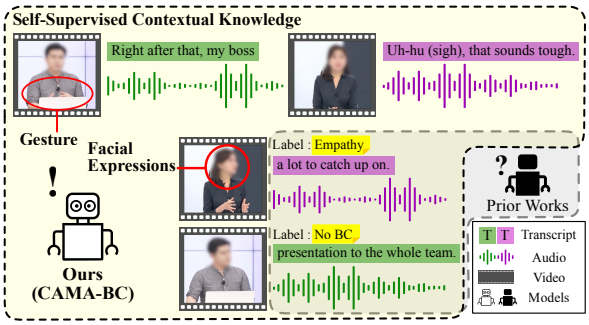


Figure 1: Schematic of the proposed data utilization in Context-Aware Multimodal Alignment for Backchannel Prediction (**CAMA-BC**), illustrating its ability to leverage a significantly broader range of information compared to the constrained scope of existing methods.

This suggests a fundamental challenge in multimodal backchannel prediction: informational asymmetry across modalities. Unlike text-based NLP tasks, where information flows sequentially, backchanneling requires understanding the subtle shift of information dominance between modalities over time. Traditional encoder-concatenation architectures, which treat all modalities equally, fail to account for temporal offsets between modalities, thereby contributing to and ignoring changes in in-

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3738–3755

	w/o video	w/ video
BPM_MT (Jang et al., 2021)	47.97	**53.77**
KoBERT+HuBERT	55.82	**56.68**

Table 1: Effect of the video modality adoption on F1 score in backchannel prediction on KC-Dialog.

	Text (# words)	Audio & Video (sec)
BC Sample	257,621	121,398
Whole Dialog	528,995	195,295
Ratio (%)	**48.70**	**62.16**

Table 2: Comparison of time duration and word count between backchannel and whole dialogue datasets. BC Sample includes both NoBC and BC samples used for backchannel alignment.

formation dominance. This asymmetry manifests in different contribution patterns: Audio contains both prosodic reactive cues and temporal contextual information, text provides rich semantic context but limited reactive signals, and video offers crucial non-verbal reactive cues.

Moreover, existing methods of backchannel prediction suffer from an imbalance between context and reaction. As demonstrated in Table 2, backchannel events represent only a sparse subset of conversational data. Additionally, the inherent subjectivity and semantic complexity of backchannel categories, such as empathy, exacerbate the training objective. These challenges cannot be resolved with simple perturbations, such as data augmentation or class re-weighting. This leads the models to overfit toward immediate pre-backchannel triggers, while neglecting broader conversational dynamics that inform natural backchanneling behavior. As a result, models are limited to pattern-matching immediate signals rather than developing genuine conversational understanding.

To address these challenges of visual information integration, informational asymmetry, and context-reaction imbalance, we introduce **CAMA-BC (Context-Aware Multimodal Alignment for Backchannel Prediction)**. This novel framework learns robust representations from unlabeled full dialogues while modeling context-dependent modality dominance through adaptive cross-attention. In this work, we propose a **Multi-Layer Multimodal Alignment (MMA)**, which utilizes hierarchical cross-attention mechanisms that account for information density differences across modalities.

Recognizing that backchannel prediction requires contextual depth, which cannot be acquired from sparse, noisy annotations alone, we employ a curriculum learning approach (Elman, 1993) consisting of two stages. First, **Context Alignment (MMA-CA)** is an unsupervised stage where the model learns rich conversational dynamics from entire dialogues. This stage enables the model to understand broad conversational patterns, acting as a regularization that prevents the upcoming stage from being biased towards reactive triggers. Next, **Backchannel Alignment (MMA-BA)** is a supervised stage that specializes in learned contextual representations for precise backchannel prediction. This stage maintains contextual understanding while developing task-specific reactive precision through cross-attention-enhanced features. Our framework offers a comprehensive solution that synergistically integrates contextual understanding with reactive precision, bridging the gap between conversational comprehension and backchannel prediction in multimodal settings.

Our main contributions can be summarized as follows:

- We identify the crucial role of the visual modality in capturing reactive components for backchannel prediction and propose a scheme that effectively integrates video information.
- To model context-dependent modality dominance, we propose **Multi-Layer Multimodal Alignment (MMA)**, which aligns multimodal features across multiple layers through adaptive cross-attention with shared weights.
- We propose a novel curriculum learning approach that consists of **Context Alignment (MMA-CA)** and **Backchannel Alignment (MMA-BA)**. The former leverages whole conversations to acquire rich contextual components, while the latter focuses on backchannel-specific information while maintaining contextual understanding.
- Through extensive experiments, we demonstrate that **Context-Aware Multimodal Alignment for Backchannel Prediction (CAMA-BC)** provides a practical and robust solution for backchannel prediction.

2 Related Work

2.1 Backchannel Prediction

Typical text-based natural language processing (NLP) models have shown success in various tasks such as speech recognition (Chan et al., 2016; Xu

Backchannel Category	Description	Examples
NoBC	No backchannel signal.	-
Continuer	Short or repeated sound indicating active listening.	"neh", "yeh", "ah"
Understanding	Longer cues showing understanding.	"uhm-", "uh-", "ah-"
Empathy	Emotions of the listener, such as surprise, sympathy, or disappointment.	"hah", "ugh", "whoa"

Table 3: Descriptions and examples of the Backchannel categories used in this work.

et al., 2021; Shakhadri et al., 2025), machine translation (Sennrich and Haddow, 2016; Wei et al., 2022), sentiment analysis (Lan, 2019; Jiang et al., 2020; Raffel et al., 2020), and question-answering (Zhong et al., 2022; Chowdhery et al., 2023).

Since Kawahara et al. (2016) formalized backchannel prediction as a computational task, researchers have recognized backchannels as crucial signals for conveying empathy and agreement in dialogue (Yngve, 1970). Early approaches, such as Ortega et al. (2020), demonstrated that incorporating transcribed text and audio can significantly improve backchannel prediction. Furthermore, they utilized listener identity information to enhance prediction. Further research on BPM_MT (Jang et al., 2021) aimed to increase accuracy by employing larger models and integrating sentiment classification as an auxiliary task, providing additional context that enhanced the predictive capabilities. Recent work by Ortega et al. (2023) extended this line by incorporating listener and speaker identities, emphasizing the speaker-listener interaction.

Previous studies have predominantly relied on audio-text combinations, systematically excluding non-verbal signals, such as visual information, despite its documented importance in human communication (Krauss et al., 1996; Scherer, 2013). Additionally, simple feature concatenation fails to account for the differential temporal dynamics of linguistic and non-linguistic signals, treating all modalities as if they contribute equally at every timestep. This narrow focus often neglects the broader conversational context that informs natural backchanneling behavior. Our approach addresses these limitations through context-aware, multimodal alignment, which captures both long-term conversational dynamics and immediate reactive cues across text, audio, and visual modalities.

2.2 Multimodal Alignment

Multimodal alignment seeks to integrate heterogeneous data sources, including text, audio, and video, into a unified representation space, thereby enabling a more nuanced understanding of semantics and facilitating cross-modal interactions. As interest in multimodal research continues to rise, remarkable progress (Wang et al., 2020; Rouditchenko et al., 2021; Sun et al., 2021; Praveen et al., 2022; Huang et al., 2022; Li et al., 2022; Shvetsova et al., 2022; Wang et al., 2022; Sadoughi et al., 2023; Girdhar et al., 2023; He et al., 2023; Zhou et al., 2024; Zhu et al., 2024) has emerged. However, many works still struggle to align more than two modalities, failing to adequately handle the geometric increase in complexity that accompanies the combination of three or more modalities.

Rouditchenko et al. (2021) proposed a self-supervised framework that aligns audio and raw video inputs in a joint embedding space without text annotations. Shvetsova et al. (2022) proposed a multimodal fusion transformer robust to the modalities and lengths, using combinatorial contrastive loss. ImageBind (Girdhar et al., 2023) focused on images, and LanguageBind (Zhu et al., 2024) centered on language, exploring efficiency by aligning one modality as an anchor. Vaswani et al. (2017) facilitated alignment using cross-attention mechanisms, where the modality with dominant performance serves as both the key and value.

These methods have demonstrated remarkable advances; however, conversational alignment remains underexplored. Unlike static contexts, conversations involve dynamically varying temporal cues and uneven information density, posing unique alignment challenges. In our approach, we depart from the static alignment assumption, ensuring alignment between each modality and its combinations. We enforce weight-sharing across the higher encoder layers to decouple the roles of encoding and cross-modal alignment, thereby encoding consistent spatiotemporal dependencies across modalities. Additionally, a two-stage curriculum learning framework is proposed that first captures broad conversational patterns before specializing in backchannel prediction.

3 Method

3.1 Problem Formulation

Given a conversation, we consider three synchronous modalities: audio, transcript, and newly considered video data. The audio dataset $\mathcal{D}_{\mathrm{a}}$ consists of amplitude values sampled at rate s_{a}, and the

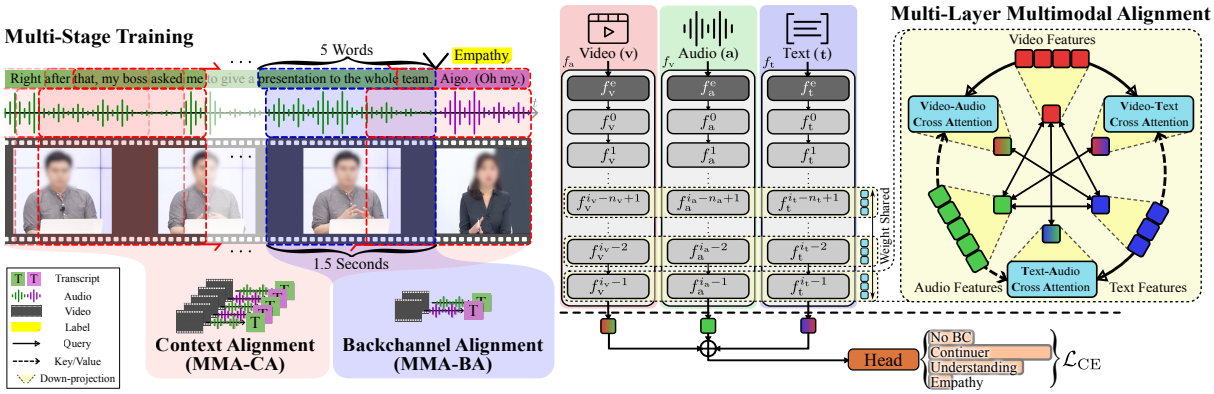


Figure 2: Illustration of the data construction and architecture of Context-Aware Multimodal Alignment for Backchannel Prediction (CAMA-BC), highlighting the Multi-Layer Multimodal Alignment (MMA) framework. The data construction process demonstrates the broader scope of the Context Alignment (MMA-CA).

transcript dataset $\mathcal{D}_\mathrm{t}$ is an ordered set of the words with their corresponding timestamps as follows:

$$\mathcal{D}_\mathrm{a} = \{a_i \mid 0 \le i < s_\mathrm{a} l_d\}, \quad (1)$$
$$\mathcal{D}_\mathrm{t} = \{(t_i, p_i) \mid 0 \le i < n_\mathrm{t}\}, \quad (2)$$

where $a_i \in \mathbb{R}$ represents audio amplitude, $t_i \in \mathbb{R}$ indicates the timestamp of each word, $p_i \in \mathcal{V}$ denotes a word from vocabulary $\mathcal{V}$, l_d is the conversation duration, and n_t is the total number of words. For the first time, in this work, we introduce video data to the backchannel prediction task, enabling the model to capture non-verbal cues in dialogue:

$$\mathcal{D}_\mathrm{v} = \{\mathbf{v}_i \mid 0 \le i < s_\mathrm{v} l_d\}, \quad (3)$$

where $\mathbf{v}_i \in \mathbb{R}^{c \times w \times h}$ represents a video frame with c channels and spatial dimensions $w \times h$, sampled at frame rate s_v.

The dataset includes annotations for listener engagement events, where each event is characterized by its timestamp $t_i^{\mathrm{BC}} \in \mathbb{R}$, vocabulary $p_i^{\mathrm{BC}} \in \mathcal{V}$ of the transcribed conversation, and category y_i from k possible backchannel classes $\mathcal{Y} = \{y_1, \ldots, y_k\}$:

$$(t_i^{\mathrm{BC}}, p_i^{\mathrm{BC}}, y_i) \in \mathcal{D}_{\mathrm{BC}}, \quad (4)$$

where $0 < i \le |\mathcal{D}_{\mathrm{BC}}|$. The detailed meaning of each category is demonstrated in Table 3. For each backchannel event $(t_i^{\mathrm{BC}}, p_i^{\mathrm{BC}}, y_i) \in \mathcal{D}_{\mathrm{BC}}$, we extract context from the preceding n seconds for the audio and video, and m words for the text:

$$\mathcal{W}_{\mathrm{a},i}^{\mathrm{BC}} = \left\{a_j \mid \left(t_i^{\mathrm{BC}} - n\right) s_a \le j < t_i^{\mathrm{BC}} s_a\right\}, \quad (5)$$
$$\mathcal{W}_{\mathrm{t},i}^{\mathrm{BC}} = \{p_j \mid i - m \le j < i\}, \quad (6)$$
$$\mathcal{W}_{\mathrm{v},i}^{\mathrm{BC}} = \left\{v_{\sigma(k)} \mid k \in \{1, 2, \ldots, l\}\right\}, \quad (7)$$

where $\sigma(k) = \left\lfloor (t_i^{\mathrm{BC}} - n) s_v + (k-1)\frac{n s_v}{l-1} \right\rfloor$ uniformly samples l frames over the interval $[t_i^{BC} - n,\ t_i^{BC})$, excluding the frame at the backchannel onset t_i^{BC}. The extracted windows are then processed into fixed-size tensors:

$$\mathbf{a}_i = \phi(\mathcal{W}_{\mathrm{a},i}^{\mathrm{BC}}) \in \mathbb{R}^{b \times n s_a}, \quad (8)$$
$$\mathbf{t}_i = \tau(\mathcal{W}_{\mathrm{t},i}^{\mathrm{BC}}) \in \mathbb{N}^{b \times m_{\mathrm{pad}}}, \quad (9)$$
$$\mathbf{v}_i = \phi(\mathcal{W}_{\mathrm{v},i}^{\mathrm{BC}}) \in \mathbb{R}^{b \times l \times c \times w \times h}, \quad (10)$$

where $\phi(\cdot)$ stacks tensors from a set, $\tau(\cdot)$ tokenizes text to integers with padding length m_{pad}, and b denotes the batch size. The final training batch $\mathcal{B}$, also called backchannel alignment dataset, combines these tensors:

$$\mathcal{B}_i = \{\mathbf{a}_i, \mathbf{t}_i, \mathbf{v}_i, y_i\}. \quad (11)$$

In practical implementation, it provides only the utterances of one speaker, with the utterances of the other speaker masked, for a realistic task configuration. The masking strategy is used as a preprocessing step to reduce input ambiguity caused by speaker alternation and overlap, rather than to encode speaker identity. We do not enforce a speaker-disjoint split, because the task focuses on modeling listener reactions rather than speaker behavior.

3.2 Context Alignment Dataset

As illustrated in Figure 1 and Table 2, backchannel responses constitute only a subset of the total conversational data, with a severe class imbalance for semantic categories such as empathy. This sparsity creates two fundamental challenges: (1) supervised learning on backchannel-only data provides insufficient examples for acquiring a robust pattern of

Model	NoBC	BC			Macro F1
		Continuer	Understanding	Empathy	
HuBERT (audio)	85.21$_{\pm 0.31}$	65.77$_{\pm 1.40}$	49.14$_{\pm 0.46}$	17.05$_{\pm 1.11}$	54.29$_{\pm 0.11}$
KoBERT (text)	84.09$_{\pm 0.74}$	57.90$_{\pm 1.94}$	30.83$_{\pm 0.65}$	17.45$_{\pm 1.07}$	47.57 $_{\pm 0.66}$
VideoMAE (video)	74.91$_{\pm 0.32}$	53.98$_{\pm 1.10}$	35.16$_{\pm 0.74}$	0.00$_{\pm 0.00}$	41.01$_{\pm 0.09}$

Table 4: F1 scores on KC-Dialog dataset with a single-modality encoder, to evaluate the impact of each modality on the backchannel prediction task.

the conversation, (2) models trained exclusively on pre-backchannel segments exhibit reactive bias, focusing solely on immediate triggers rather than understanding broader conversational flow.

To address these limitations, we propose Context Alignment (CA), an unsupervised pre-training phase that leverages the complete conversational corpus. For each word-timestamp pair $(t_i^{\mathrm{CA}}, p_i^{\mathrm{CA}}) \in \mathcal{D}_{\mathrm{CA}}$, we extract context windows of identical dimensions to those used in backchannel prediction. By maintaining identical window dimensions across two phases, we ensure that contextual representations learned during unsupervised pre-training serve as initialization for fine-tuning the model towards backchannel prediction. In contrast, the model learns to distinguish between general conversational patterns and backchannel-specific triggers. By exposing the model to temporally synchronized multimodal context beyond backchannel events, the pre-training encourages temporal alignment across modalities.

3.3 Multi-Layer Multimodal Alignment

Unlike the backchannel alignment dataset, the context data constructed in Section 3.2 does not contain explicit labels. Therefore, we need an approach to learning contextual information from unlabeled dialogue data. Our preliminary analysis (Table 4) reveals asymmetric information density across modalities for backchannel prediction. Audio shows the highest predictive power, containing both reactive and contextual information. The text is followed by providing rich contextual semantics, but with limited reactive cues. In contrast, video contains sparse linguistic content but contributes crucial non-verbal reactive signals.

This asymmetry motivates a hierarchical alignment strategy where information-dense modalities assist information-sparse ones, rather than treating all modalities equivalently. It determines explicitly which modality acts as the Query and which as the Key/Value in cross-modal interactions, prioritizing them in the order of audio, text, and video through cross-modal feature alignment.

As shown in Figure 2, we distinguish between embedding layers (f_{a}^e, f_{t}^e, and f_{v}^e) and encoder layers ($\{f_{\mathrm{a}}^l \mid 0 \le l < i_{\mathrm{a}}\}$, $\{f_{\mathrm{t}}^l \mid 0 \le l < i_{\mathrm{t}}\}$, and $\{f_{\mathrm{v}}^l \mid 0 \le l < i_{\mathrm{v}}\}$) for audio, text, and video models, respectively. The feature extraction process can be expressed as:

$$\begin{aligned}
\mathbf{a}_i^0 &= f_{\mathrm{a}}^e(\mathbf{a}_i), \;\; \mathbf{a}_i^{l+1} = f_{\mathrm{a}}^l(\mathbf{a}_i^l), \\
\mathbf{t}_i^0 &= f_{\mathrm{t}}^e(\mathbf{t}_i), \;\; \mathbf{t}_i^{l+1} = f_{\mathrm{t}}^l(\mathbf{t}_i^l), \\
\mathbf{v}_i^0 &= f_{\mathrm{v}}^e(\mathbf{v}_i), \;\; \mathbf{v}_i^{l+1} = f_{\mathrm{v}}^l(\mathbf{v}_i^l).
\end{aligned} \tag{12}$$

We align the features in the last n_l layers of each modality encoder as follows:

$$\begin{aligned}
\mathcal{J} = \{ j = (j_{\mathrm{a}}, j_{\mathrm{t}}, j_{\mathrm{v}}) \mid \\
i_{\mathrm{a}} - n_{\mathrm{a}} + 1 \le j_{\mathrm{a}} < i_{\mathrm{a}}, \\
i_{\mathrm{t}} - n_{\mathrm{t}} \; + 1 \le j_{\mathrm{t}} < i_{\mathrm{t}}, \\
i_{\mathrm{v}} - n_{\mathrm{v}} + 1 \le j_{\mathrm{v}} < i_{\mathrm{v}} \} ,
\end{aligned} \tag{13}$$

for audio, text, and video, respectively, with the same number of selections $n_{\mathrm{a}} = n_{\mathrm{t}} = n_{\mathrm{v}} = n_l$. This strategic choice emphasizes semantic alignment rather than modality-specific features. For the alignment at the $j = (j_{\mathrm{a}}, j_{\mathrm{t}}, j_{\mathrm{v}})$, we first project the sequence-level features into a shared space and then average-pool the projected features into sample-level embeddings. The model aligns embeddings from the same sample while treating the remaining samples in the batch as negatives:

$$\mathcal{L}_{\mathrm{MMA}}^j = \sum_{\alpha \in \mathcal{S}_j, \beta \in \mathcal{S}_j \setminus \{\alpha\}} \delta(\alpha, \beta), \tag{14}$$

$$\begin{aligned}
\delta(\mathbf{v}, \mathbf{w}) &= \mathrm{diag}\left(\mathrm{LogSoftmax}\left(\frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\| \, \|\mathbf{w}\|} \right) \right), \\
\mathcal{S}_j &= \left\{ \bar{\mathbf{a}}^{j_{\mathrm{a}}}, \bar{\mathbf{t}}^{j_{\mathrm{t}}}, \bar{\mathbf{v}}^{j_{\mathrm{v}}} \right\},
\end{aligned} \tag{15}$$

where $\bar{\mathbf{m}}^{j_{\mathrm{m}}} = \mathrm{AvgPool}\big(W_{\mathrm{m}} \mathbf{m}^{j_{\mathrm{m}}}\big)$ denotes the sample-level embedding of modality $\mathrm{m} \in \{\mathrm{a}, \mathrm{t}, \mathrm{v}\}$, and $W_{\mathrm{a}} \in \mathbb{R}^{d_{\mathrm{a}} \times d}$, $W_{\mathrm{t}} \in \mathbb{R}^{d_{\mathrm{t}} \times d}$, and $W_{\mathrm{v}} \in \mathbb{R}^{d_{\mathrm{v}} \times d}$ are linear transformations that project encoder features into a shared d-dimensional space. Our empirical analysis shows that applying alignment to earlier layers proves counterproductive. These layers lack the necessary abstraction for effective cross-modal fusion, primarily capturing modality-specific features.

However, the representations across video, audio, and text modalities exhibit significant differences, making direct alignment using simple alignment difficult. To address this issue, we employ a hierarchical cross-attention mechanism, motivated

by our experimental findings indicating varying information densities (e.g., audio > text > video) and distributions across modalities, as shown in Table 4. The cross-attention layer enables a query modality to be refined by selectively attending to information from an assistant key/value modality with denser information. To incorporate attention-guided enhancement, we extend the alignment loss with average-pooled cross-attended representations as follows:

$$\mathcal{L}_{\text{MMA}}^{j,*} = \mathcal{L}_{\text{MMA}}^{j} + \delta\left(\bar{\mathbf{a}}^{j_\text{a}}, \bar{\mathbf{c}}_{\text{vt}}^{j}\right) + \delta\left(\bar{\mathbf{t}}^{j_\text{t}}, \bar{\mathbf{c}}_{\text{va}}^{j}\right) + \delta\left(\bar{\mathbf{v}}^{j_\text{v}}, \bar{\mathbf{c}}_{\text{ta}}^{j}\right), \quad (16)$$

where $\bar{\mathbf{c}}_{\text{vt}}^{j}$, $\bar{\mathbf{c}}_{\text{va}}^{j}$, and $\bar{\mathbf{c}}_{\text{ta}}^{j}$ denote the average-pooled cross-attended representations computed from video-text, video-audio, and text-audio feature pairs, respectively. The hierarchy ensures that information flows from denser key/value modalities to sparser query modalities. For instance, in $\delta\left(\bar{\mathbf{a}}^{j_\text{a}}, \bar{\mathbf{c}}_{\text{vt}}^{j}\right)$, the video feature $\bar{\mathbf{v}}^{j_\text{v}}$ is used as the query and the relatively informative text feature $\bar{\mathbf{t}}^{j_\text{t}}$ is used as the key and value, producing a representation aligned with the audio feature $\bar{\mathbf{a}}^{j_\text{a}}$.

Traditional layer-specific attention would learn modality-specific transformations at each layer, potentially leading to overfitting to layer-dependent features and competing for the same knowledge space as the encoders. Instead, we share cross-attention weights across all layers in $\mathcal{J}$. This strategy encourages the cross-attention layer to learn consistent cross-modal interaction patterns and a modality-agnostic approach to identifying salient inter-token relationships. This approach facilitates smoother learning by robustly pinpointing crucial tokens across different representational levels.

This learned attention pattern is then directly utilized in the final prediction stage as follows:

$$\mathbf{y}_{\text{pred}} = W_h \cdot \text{cat}\left\{\bar{\mathbf{a}}^{j_\text{a}}, \bar{\mathbf{c}}_{\text{va}}^{j}, \bar{\mathbf{c}}_{\text{ta}}^{j}\right\} + \mathbf{b}_h, \quad (17)$$

where $\mathbf{y}_{\text{pred}}$ denotes the logits, $\text{cat}\left(\cdot, \cdot, \dots\right)$ is concatenation of the given tensors, and $W_h \in \mathbb{R}^{3d \times k}$ and $\mathbf{b}_h \in \mathbb{R}^k$ are weight and bias in the classification head, respectively. While the audio serves as the primary anchor, video and text are enhanced using dense information from the audio. This architecture preserves the information hierarchy while enabling cross-modal enhancement.

3.4 Multi-Stage Training

We construct our Multi-Layer Multimodal Alignment (MMA) framework through a two-phase training process: Context Alignment (MMA-CA) followed by Backchannel Alignment (MMA-BA). The first phase (Section 3.4.1) trains the model on general conversational context and cross-modal relationships. In contrast, the second phase (Section 3.4.2) adapts the model specifically for backchannel prediction while preserving contextual knowledge.

3.4.1 Context Alignment

The Context Alignment phase utilizes the data in Section 3.2 to train the model on broader conversational dynamics and inter-modal correlations. During this phase, we optimize solely using the MMA loss defined in Equation 16:

$$\mathcal{L}_{\text{CA}} = \sum_{j \in \mathcal{J}} \mathcal{L}_{\text{MMA}}^{j,*}, \quad (18)$$

where $\mathcal{J}$ represents the layers selected for alignment. The model learns to recognize correlations between verbal, vocal, and visual elements of natural dialogue by training on continuous conversation segments of equivalent length to the backchannel-annotated samples. This equivalence in segment length ensures that the learned representations can be directly applied to the downstream task.

3.4.2 Backchannel Alignment

The Backchannel Alignment phase fine-tunes the model on backchannel-annotated data while maintaining the contextual knowledge learned in the first phase. We optimize a combined objective that incorporates both classification and alignment losses:

$$\mathcal{L}_{\text{BA}} = \mathcal{L}_{\text{CE}}\left(\mathbf{y}_{\text{pred}}, y\right) + \lambda \sum_{j \in \mathcal{J}} \mathcal{L}_{\text{MMA}}^{j,*}, \quad (19)$$

where $\mathcal{L}_{\text{CE}}\left(\cdot, \cdot\right)$ denotes the cross-entropy loss for backchannel classification, and λ controls the contribution of the alignment loss (set to 0.1 in our experiments). This dual objective serves two purposes: (1) it preserves contextual knowledge during task-specific adaptation, and (2) it strengthens the cross-modal relationships relevant to backchannel prediction.

4 Experiments

4.1 Experimental Setup

Datasets. We utilize three datasets: the Korean Counseling Dialog Dataset (KC-Dialog), the

Dataset	NoBC	BC			Train	Validation	Test	Total	#Samples for MMA-CA
		Continuer	Understanding	Empathy					
KC-Dialog	34,710	19,635	9,494	5,581	55,536	3,470	10,414	69,420	130,278
BACKSpeech	4,907	4,046	352	509	7,852	490	1,472	9,814	34,580

Table 5: Distribution of each backchannel category and the number of samples used for MMA-CA. Following prior works, we ensured that the NoBC and BC samples had equal numbers. The dataset was then split into training, validation, and test sets in an $8 : 0.5 : 1.5$ ratio.

Backchannel Annotation Corpus in Korean Speech Dataset (BACKSpeech) (ETRI, 2023), and the Switchboard (SWBD) corpus (Godfrey et al., 1992). KC-Dialog consists of counselor-client videos, each about 50 minutes long, with transcripts annotated for backchannel classes. BACKSpeech contains 55 videos totaling 20 hours from diverse media sources, including radio, news, and debates. While both datasets provide backchannel-annotated transcripts, BACKSpeech features more dynamic visual content with varying scenes and visual materials, in contrast to speaker-centric recordings of KC-Dialog. The audio is sampled at $s_a = 16000$Hz and video is sampled at $s_v = 30$Hz. A description of SWBD is provided in Section C.1.

Following the categorization scheme of BPM_MT (Jang et al., 2021), we classify backchannels into four classes: NoBC, Continuer, Understanding, and Empathy, with examples and descriptions provided in Table 3, and class distributions detailed in Table 5. For feature extraction, we analyze temporal windows of $n = 1500$ms and $m = 5$ words following each backchannel onset, adhering to established protocols from previous studies (Ortega et al., 2020; Jang et al., 2021). Corresponding to the temporal segment of the video, we uniformly sample $l = 12$ frames following Tong et al. (2022).

Evaluation. We evaluate performance using the Macro F1 score, which is the average F1 score across all classes. We report the best performance over 60 epochs with early stopping. All results are averaged over three random seeds with statistical significance testing.

Baselines. We compare our method with several baseline models utilizing different modality combinations of text, audio, and video. As text-only baselines, we use KoBERT (SKTBrain, 2021) and two large language models, GPT-4.1 (OpenAI, 2025) and Llama 4.0 Scout (Meta, 2025). We adopt Ortega (Ortega et al., 2020) and BPM_MT (Jang et al., 2021), prior works on backchannel prediction that utilize both text and audio, as baselines in this work. To examine whether visual information aids even in naive integration, we extend the existing Ortega and BPM_MT models by incorporating a pre-trained VideoMAE (Tong et al., 2022) via simple concatenation, resulting in Ortega-V and BPM-V. We also conduct experiments on GPT-4.1 (OpenAI, 2025) and Llama 4.0 Scout (Meta, 2025) using text-only inputs, and on Gemini-2.5-Flash (Google Deepmind, 2025) for multimodal inputs, to verify the performance of backchannel prediction compared to well-known large language models.

Implementation Details. Our model architecture incorporates pre-trained models specialized for each modality. For text processing, we utilize KoBERT (SKTBrain, 2021), a Korean-specific variant of BERT (Devlin et al., 2019), to effectively handle Korean language content. Audio feature extraction utilizes HuBERT (Hsu et al., 2021), chosen for its self-supervised speech representations. For video, we use VideoMAE (Tong et al., 2022), which is trained on Kinetics (Kay et al., 2017), a dataset covering diverse human actions relevant to our task.

For the Ortega and BPM baselines, we follow the original feature settings reported in the respective papers, including their audio representations.

4.2 Experimental Results

Table 6 demonstrates that CAMA-BC consistently outperforms baseline models across different architectures, achieving robust improvements in minority classes such as Continuer and Empathy, where contextual understanding is most critical. A key finding emerges from comparing integration approaches. While direct visual feature concatenation (BPM-V) yields only modest improvement, as shown in Table 1, hierarchical alignment of the CAMA-BC achieves substantial gains. This demonstrates that the value of visual information emerges through proper cross-modal alignment rather than naive feature combination.

Both large language models using text-only and multimodal inputs showed limited performance, confirming that backchannel prediction fundamentally requires multimodal reactive signals. We

Dataset	Model	Modality	NoBC	BC			Macro F1
				Continuer	Understanding	Empathy	
KC-Dialog	KoBERT	T	84.09±0.74	57.90±1.94	30.83±0.65	17.45± 1.07	47.57 ±0.66
	Llama 4.0 Scout	T	53.02±0.61	25.07±1.21	12.25±0.86	4.48±0.20	23.71±0.20
	GPT-4.1	T	57.39±0.30	33.03±0.47	9.86±0.09	6.73±0.06	26.75±0.17
	Ortega	A,T	84.53±0.51	60.54±1.20	24.25±0.98	7.59± 1.33	44.23 ±0.83
	BPM_MT	A,T	85.73±0.27	61.55±1.72	30.43±2.77	14.18± 2.33	47.97 ±0.78
	Gemini-2.5-Flash	A,T,V	44.63±1.11	6.02±1.24	3.62±0.48	4.68±0.89	14.74±0.52
	Ortega-V	A,T,V	84.46±0.39	65.16±1.18	42.74±1.00	7.25±0.18	49.90±0.27
	BPM-V	A,T,V	86.96±0.18	65.72±1.49	45.50±1.97	16.89±2.09	53.77±0.09
	CAMA-BC (Ours)	A,T,V	**89.90**±0.07	**72.73**±0.13	**49.99**±0.25	**21.49**±0.49	**58.53**±0.07
BACKSpeech	KoBERT	T	67.04±0.65	62.80±0.21	0.00±0.00	0.00± 0.00	32.46 ±0.21
	Llama 4.0 Scout	T	18.69±0.99	27.93±0.26	4.63±0.49	2.11±0.39	13.34±0.19
	GPT-4.1	T	30.21±0.39	43.73±0.74	2.51±0.36	3.12±0.13	19.89±0.33
	Ortega	A,T	68.21±0.60	61.39±0.44	0.00±0.00	0.00± 0.00	32.40 ±0.15
	BPM_MT	A,T	69.10±1.03	62.66±0.74	0.00±0.00	0.00± 0.00	32.94 ±0.31
	Gemini-2.5-Flash	A,T,V	26.19±2.42	11.64±1.22	1.51±1.16	3.65±0.64	10.75±0.39
	Ortega-V	A,T,V	69.24±0.68	62.63±0.33	0.00±0.00	1.22± 1.06	33.27 ±0.02
	BPM-V	A,T,V	70.78±2.79	63.26±0.58	0.00±0.00	1.37± 2.37	33.85 ±0.17
	CAMA-BC (Ours)	A,T,V	**74.22**±0.42	**67.81**±0.65	**8.48**±1.80	**6.28**±2.09	**39.20** ±0.14

Table 6: Evaluation of backchannel prediction on the KC-Dialog dataset and BACKSpeech dataset. V† refers to the image sequence used as a vision input.

Model	NoBC	BC			Macro F1
		Continuer	Understanding	Empathy	
CAMA-BC w/o MMA	88.94	70.82	48.75	18.10	56.68
CAMA-BC w/o MMA-BA	88.98	69.71	49.64	17.96	56.57
CAMA-BC w/o MMA-CA	89.73	72.15	48.35	20.13	57.59
CAMA-BC (Ours)	**89.90**	**72.73**	**49.99**	**21.49**	**58.53**

Table 7: Evaluation on the impact of MMA-BA and MMA-CA on CAMA-BC on the KC-Dialog.

Model	Modality	NoBC	BC	Macro F1
Llama 4.0 Scout	T	28.80±0.61	50.21±0.34	39.50±0.45
GPT-4.1	T	44.15±0.25	42.61±0.41	43.38±0.29
Ortega	A,T	68.27±0.27	71.07±1.04	69.67±0.65
BPM_ST	A,T	70.41±0.52	77.55±0.45	73.98±0.45
Gemini-2.5-Flash	A,T,V	67.24±0.40	7.11±0.74	37.18±0.30
Ortega-V	A,T,V	70.27±0.33	72.07±0.80	71.17±0.55
BPM-V	A,T,V	72.31±0.48	78.34±0.37	75.33±0.39
CAMA-BC (Ours)	A,T,V	**76.24**±0.44	**79.70**±0.42	**77.97**±0.38

Table 8: Evaluation on the performance of CAMA-BC on the SWBD Dataset with generated videos.

observed substantial prompt sensitivity for the LLM/VLM baselines; however, despite exploring multiple prompt variants, the overall performance trend remained unchanged.

4.3 Analysis

4.3.1 Ablation Study

The detailed performance analysis in Table 6 reveals that CAMA-BC with MMA components achieves powerful improvements in Continuer and Empathy classes, which represent minority classes in our dataset (see Table 5). Table 7 shows the effect of MMA-CA and MMA-BA. The performance benefit persists even without MMA-BA, though with some degradation due to feature misalignment during transfer learning.

Interestingly, without MMA-BA, the model achieved only a moderate improvement in classification, resulting in the loss of reactive information and corruption of the representation space during the transfer process. This suggests that enhancing the representation does not necessarily lead to improved classification. On the other hand, without MMA-CA, the model achieves a better classification result because the enhanced knowledge directly aids the classification. Our framework integrates both contextual understanding and reactive precision through MMA-CA and MMA-BA components. The results show that the alignment loss serves as an unsupervised objective during Context Alignment and as a regularization during Backchannel Alignment, preventing catastrophic forgetting on transfer to classification.

4.3.2 SWBD Dataset

We experiment on the SWBD dataset to verify the effectiveness of CAMA-BC with the English dataset. Since the original SWBD dataset lacks a visual modality, we generated video clips aligned with the corresponding audio segments, as described in Section C.2. The experimental results are presented in Table 8. Even with generated videos, visual modality consistently improves performance across all baselines. CAMA-BC achieves the best overall performance among all models, demonstrating cross-linguistic generalizability. These results suggest our approach captures the fundamental conversational dynamics rather than language-specific patterns.

5 Conclusion

We identified three critical limitations in natural interactive conversational AI: the systematic exclusion of visual information, temporal misalignment between modalities, and context-reaction imbalance. Our Context-Aware Multimodal Alignment for Backchannel Prediction (CAMA-BC) represents the first systematic integration of visual information into backchannel prediction. Through hierarchical cross-modal alignment, which addresses information asymmetry across modalities, CAMA-BC incorporates the visual modality without falling into the limitations of simple feature concatenation, which often fails to leverage the full potential of visual signals. Comprehensive experiments across Korean (KC-Dialog, BACKSpeech) and English (SWBD) datasets demonstrated consistent improvements, suggesting that hierarchical cross-modal alignment not only enhances class-level prediction fidelity but also provides a language-agnostic framework that robustly captures universal conversational cues.

Limitations

While our approach yields promising results, several limitations warrant discussion. Although our modality hierarchy was empirically validated, it may not generalize well to different types of interaction. This suggests the need for a dynamic hierarchy detection mechanism that adapts to varying conversational contexts. Following prior work, we use a 1500ms audio window and a 5-word text context as a default. The additional experiment in Table A.5 demonstrates that this may not fully capture long-range conversational dependencies. Future work could explore adaptive or hierarchical management of temporal contexts. Moreover, integrating video introduces latency and increased inference cost, potentially limiting real-time applicability; however, as detailed in Section D, it is still affordable for real-time applications. Finally, backchannel categorization inherently involves subjective properties. Our model inherits backchannel categories, such as "Empathy" or "Understanding," which are inherently subjective. Table A.6 demonstrates the robustness on the noisy labels, but it still implicitly inherits these biases without explicitly accounting for annotation uncertainty or inter-annotator disagreement.

Acknowledgments

This research was supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25480253), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), RS-2024-00457882, AI Research Hub Project, and RS-2024-00509257, Global AI Frontier Lab).

References

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP*.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. 2025. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2403–2410.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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

Jeffrey L. Elman. 1993. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99.

ETRI. 2023. Backchannel annotation corpus in korean speech. https://github.com/etri/etri-miai.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Google Deepmind. 2025. Gemini-2.5. https://blog.google/technology/google-deepmind/gemini-2-5-native-audio.

Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14867–14878.

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. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Yi Huang, Xiaoshan Yang, Ji Zhang, and Changsheng Xu. 2022. Relative alignment network for source-free multimodal video domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1652–1660, New York, NY, USA. Association for Computing Machinery.

Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. BPM_MT: Enhanced backchannel prediction model using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3447–3452, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems. In *proceedings of INTERSPEECH 2016*, pages 2890–2894.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Robert M Krauss, Yihsiu Chen, and Purnima Chawla. 1996. Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In *Advances in experimental social psychology*, volume 28, pages 389–450. Elsevier.

Zhenzhong Lan. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Peiguang Li, Xian Sun, Hongfeng Yu, Yu Tian, Fanglong Yao, and Guangluan Xu. 2022. Entity-oriented multi-modal alignment and fusion network for fake news detection. *IEEE Transactions on Multimedia*, 24:3455–3468.

Fatik Baran Mandal. 2014. Nonverbal communication in humans. *Journal of human behavior in the social environment*, 24(4):417–421.

Meta. 2025. Llama-4.0 scout. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

OpenAI. 2025. Gpt-4.1. https://openai.com/index/gpt-4-1/.

Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, jeez! or uh-huh? a listener-aware backchannel predictor on asr transcriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8064–8068. IEEE.

Daniel Ortega, Sarina Meyer, Antje Schweitzer, and Ngoc Thang Vu. 2023. Modeling speaker-listener interaction for backchannel prediction. *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L Koerich, Simon Bacon, Patrick Cardinal, et al. 2022. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2486–2495.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. 2021. Avlnet: Learning audio-visual language representations from instructional videos. In *proceedings of INTERSPEECH 2021*, pages 1584–1588.

Najmeh Sadoughi, Xinyu Li, Avijit Vajpayee, David Fan, Bing Shuai, Hector Santos-Villalobos, Vimal Bhat, and Rohith Mv. 2023. Mega: Multimodal alignment aggregation and distillation for cinematic video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23331–23340.

Klaus R Scherer. 2013. The functions of nonverbal signs in conversation. In *The social and psychological contexts of language*, pages 225–244. Psychology Press.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91.

Syed Abdul Gaffar Shakhadri, Kruthika KR, and Kartik Basavaraj Angadi. 2025. Samba-asr state-of-the-art speech recognition leveraging structured state-space models. *arXiv preprint arXiv:2501.02832*.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20020–20029.

SKTBrain. 2021. Korean bert pre-trained cased (kobert). https://github.com/SKTBrain/KoBERT.

Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021. Multimodal cross- and self-attention network for speech emotion recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4275–4279.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qinxin Wang, Hao Tan, Sheng Shen, Michael Mahoney, and Zhewei Yao. 2020. MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2030–2038, Online. Association for Computational Linguistics.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. 2022. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12186–12195.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034. IEEE.

Victor H Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, Chicago Linguistic Society, Chicago*, pages 567–578.

Qihuang Zhong, Liang Ding, Yibing Zhan, Yu Qiao, Yonggang Wen, Li Shen, Juhua Liu, Baosheng Yu, Bo Du, Yixin Chen, et al. 2022. Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue. *arXiv preprint arXiv:2212.01853*.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17114–17122.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Symbol	Description
$\mathcal{D}_{\mathrm{a}}$	Audio dataset consisting of amplitude values sampled at rate s_a.
$\mathcal{D}_{\mathrm{t}}$	An ordered set of words with their timestamps, where p_i is a word and t_i is its timestamp.
$\mathcal{D}_{\mathrm{v}}$	Video dataset consisting of video frames sampled at rate s_v.
$\mathcal{D}_{\mathrm{BC}}$	Backchannel dataset containing annotations for listener engagement events.
D_{CA}	Context alignment dataset constructed from full dialogues for unsupervised context learning.
$\mathcal{W}_{\mathrm{a}}^{\mathrm{BC}}, \mathcal{W}_{\mathrm{t}}^{\mathrm{BC}}, \mathcal{W}_{\mathrm{v}}^{\mathrm{BC}}$	Context windows for audio, text, and video modalities for backchannel prediction.
$\mathcal{W}_{\mathrm{a}}^{\mathrm{CA}}, \mathcal{W}_{\mathrm{t}}^{\mathrm{CA}}, \mathcal{W}_{\mathrm{v}}^{\mathrm{CA}}$	Context windows for audio, text, and video modalities for context alignment.
$\mathbf{a}_i, \mathbf{t}_i, \mathbf{v}_i$	Audio, text, and video tensor for the i-th context window.
$\mathcal{B}_i$	i-th batch of backchannel data, containing audio, text, and video tensors.
$\mathcal{J}$	Set of selected layer groups $j = (j_a, j_t, j_v)$ used for multimodal alignment.
$\bar{\mathbf{a}}^{j_{\mathrm{a}}}, \bar{\mathbf{t}}^{j_{\mathrm{t}}}, \bar{\mathbf{v}}^{j_{\mathrm{v}}}$	Average-pooled sample-level embeddings at selected layers j_a, j_t, j_v for audio, text, and video.
$\bar{\mathbf{c}}_{vt}^{j}, \bar{\mathbf{c}}_{va}^{j}, \bar{\mathbf{c}}_{ta}^{j}$	Average-pooled cross-attended representations of modality pairs at layer group j.
$\mathcal{S}_j$	Set of average-pooled modality embeddings used for alignment at layer group j.
$W_{\mathrm{a}}, W_{\mathrm{t}}, W_{\mathrm{v}}$	Linear transformations for audio, text, and video features.
$\mathcal{L}_{\mathrm{MMA}}^{j}$	Multimodal Alignment loss at layer j.
$\mathcal{L}_{\mathrm{MMA}}^{j,*}$	Multimodal Alignment loss with the cross-attention mechanism at layer group j.
$\mathcal{L}_{\mathrm{CE}}$	Cross-entropy loss for backchannel classification.
$\delta(\mathbf{v}, \mathbf{w})$	Similarity function for aligning features $\mathbf{v}$ and $\mathbf{w}$.
$\mathbf{y}_{\mathrm{pred}}$	Predicted logits for backchannel classification.
W_h	Weight matrix for the final classification head.
$\mathbf{b}_h$	Bias vector for the final classification head.
λ	Hyperparameter controlling the contribution of the alignment loss.
$i_{\mathrm{a}}, i_{\mathrm{t}}, i_{\mathrm{v}}$	Number of layers in audio, text, and video encoders.
n_l	Number of layers used for alignment in the Multimodal Alignment framework.

Table A.1: Description for the mathematical symbols used in the paper.

A Mathematical Notations

Table A.1 provides a comprehensive reference for the mathematical symbols and notations used throughout the paper.

B Additional Experimental Results

B.1 Ablations for Hyperparameters

We conduct experiments to evaluate the effectiveness of the components that comprise our method. First, in Table A.2, we see the results of adjusting the lambda scale constant in our method. Overall, we observe that it performs nearly as well as if it is not too large, so it does not violate the classification entropy loss. In Table A.3, we also observed that selecting more layers towards the front of the model did not guarantee better performance when aligning layers. This is likely because, as discussed in Section 3.3, the early, abstracted information hinders alignment.

B.2 Ablation for Alignment Methods

We checked the performance impact on the backchannel prediction task for different multimodality alignment-seeking methodologies. In many cases, techniques such as cross-attention are designed to use two alignment targets, which creates structural problems when applied to our scenario, where three modality alignments are required. As shown in Table A.4, when we expand these methodologies, they perform relatively poorly, demonstrating the importance of the assistance and selectivity between modalities that we sought.

λ	NoBC	BC			Macro F1
		Continuer	Understanding	Empathy	
1.0	89.57	71.75	46.63	14.87	55.71
0.5	89.68	72.68	47.47	15.80	56.41
0.2	90.02	72.81	49.07	18.98	57.72
0.1 (Ours)	89.90	72.73	49.99	21.49	**58.53**
0.05	89.68	71.71	47.60	21.74	57.68
0	88.98	69.71	49.64	17.96	56.57

Table A.2: Ablation study on the loss weight λ on the KC-Dialog dataset.

n_l	Layer	NoBC	BC			Macro F1
			Continuer	Understanding	Empathy	
11	0-11	90.07	73.07	48.11	21.16	58.10
6	6-11	90.05	73.00	49.63	20.21	58.22
3	9-11	89.90	72.73	49.99	21.49	**58.53**
1	11	89.73	72.11	48.40	22.59	58.21

Table A.3: Analysis of the impact of alignment depth in Multi-Layer Multimodal Alignment (MMA) on the KC-Dialog dataset.

Alignment	NoBC	BC			Macro F1
		Continuer	Understanding	Empathy	
Praveen	88.17	69.69	49.85	19.33	56.75
Sun	89.46	71.84	43.59	22.26	56.79
MMA	89.90	72.73	49.99	21.49	**58.53**

Table A.4: Comparison of Multi-Layer Multimodal Alignment (MMA) with other alignment methods on the KC-Dialog dataset.

Model	Text Length	Audio/Video Length	NoBC	BC			Macro F1
				Continuer	Understanding	Empathy	
BPM-V	5 words	1500ms	86.96$_{\pm 0.18}$	65.72$_{\pm 1.49}$	45.50$_{\pm 1.97}$	16.89$_{\pm 2.09}$	53.77$_{\pm 0.09}$
	10 words	3000ms	88.99$_{\pm 0.37}$	68.01$_{\pm 0.33}$	46.92$_{\pm 0.76}$	17.01$_{\pm 0.15}$	55.50$_{\pm 0.30}$
CAMA-BC (Ours)	5 words	1500ms	89.90$_{\pm 0.07}$	72.73$_{\pm 0.13}$	49.99$_{\pm 0.25}$	21.49$_{\pm 0.49}$	58.53$_{\pm 0.07}$
	10 words	3000ms	91.21$_{\pm 0.37}$	73.91$_{\pm 0.23}$	51.14$_{\pm 0.22}$	22.69$_{\pm 0.12}$	59.76$_{\pm 0.11}$

Table A.5: Analysis of backchannel prediction performance with different input lengths on the KC-Dialog dataset.

Model	Ratio	NoBC	BC			Macro F1
			Continuer	Understanding	Empathy	
BPM-V	0%	86.96	65.72	45.50	16.89	53.77
	5%	86.73	64.87	0.00	0.00	37.90 (-15.87)
	10%	86.46	64.55	0.00	0.00	37.75 (-16.02)
CAMA-BC	0%	89.90	72.73	49.99	21.49	58.53
	5%	88.25	69.92	44.96	11.81	53.74 (-4.79)
	10%	87.99	69.55	43.08	10.04	52.67 (-5.86)

Table A.6: Performance of BPM-V and CAMA-BC under different ratios of synthetic label noise on the KC-Dialog dataset, where a fixed percentage of training labels was randomly flipped.

B.3 Analysis of Input Length

We conducted experiments to evaluate the impact of input length on model performance. Specifically, as shown in Table A.5, we compared a 5-word, 1500ms input with an extended 10-word, 3000ms input and observed that increasing the input length provides additional contextual cues, leading to modest performance improvements. However, our primary goal is not to maximize performance by simply expanding the input window. Instead, we focus on designing efficient and generalizable architectures that perform reliably under practical constraints, ensuring robust performance even with limited input length.

B.4 Robustness to Annotation Noise

Backchannel annotations are created according to detailed guidelines to maintain consistency; however, there is no absolute ground truth, and subjectivity is inevitably involved. To examine how such annotation disagreement and noise affect model robustness, we conducted a synthetic label noise experiment. Specifically, we randomly flipped a portion of the training labels (5% and 10%) to simulate annotator disagreement and evaluated model performance under these noisy conditions. The results, shown in Table A.6, indicate that BPM-V suffers substantial drops in Macro F1 as the noise ratio increases, especially for minority classes. In contrast, CAMA-BC shows only a modest performance decline and remains stable even when label noise is introduced. This suggests that our model is not heavily affected by label quality uncertainty and learns stable multimodal representations, maintaining robustness even under subjective labeling noise.

B.5 Qualitative Results

Figure A.1 shows prediction examples, while Figures A.2 and A.3 visualize attention patterns in successful and failed CAMA-BC cases to identify the most essential multimodal features. The visual modality analysis reveals consistent attention to early video frames, relatively capturing more pre-utterance cues than other modalities. Additionally, failure cases indicate that the audio and text are not informative enough, and the speaker is using minimal gestures and expressions. This demonstrates misaligned visual attention where the model inadequately focuses on speaker-related content. In both cases, the audio tends to focus on the silence between utterances, which is a significant indicator of interaction.

C SWBD Video Dataset

C.1 SWBD Dataset

We utilize the Switchboard (SWBD) corpus to evaluate the effectiveness of CAMA-BC in another language setting. SWBD is a large-scale English telephone speech dataset comprising over 2,400 conversations with more than 500 adult speakers discussing a wide range of topics. Each conversation lasts approximately 6–10 minutes and includes detailed utterance-level transcriptions for all audio recordings.

For text and audio feature extraction, we analyze a temporal window of $n = 1500$ms and $m = 5$ words following each backchannel onset, consistent with our previous experiments. As SWBD does not contain video, we generate video data synthetically for the same temporal segments to incorporate the visual modality.

C.2 Video Generation

C.2.1 Reference Image Generation

To address the absence of visual data in the SWBD corpus, we first generate reference images for each

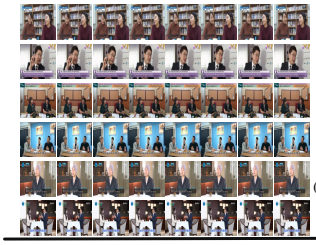
Video Frames	Speaker Transcript	Ground Truth Listener Transcript	Ground Truth Backchannel Category	CAMA-BC Prediction	BPM-V Prediction
	나누러 가야 될 것 같아요. (I think I have to go and distribute it.)	남미 (South America)	NoBC	Understanding	Continuer
	싸이클린 코인이 함께하는 게 저희의 (that the Cyclin Coin is involved)	예 (Yeah)	Understanding	NoBC	Continuer
	세차문화가 거의 이십 년 동안 (the car washing has been ... for nearly twenty years)	네 (Uh-huh)	Continuer	Understanding	NoBC
	인간의 기본적인 권리를 누릴 수 (To be able to enjoy the human rights)	음 (Um)	Continuer	Continuer	Continuer
	공부잘하고 한국에 관심 많은 학생들을 (Students who are good at studying and interested in Korea)	아아 (Ah, I see)	Understanding	Understanding	Continuer
	이장인 저희 아버지께서 하셔야 되는 (As a village chief, my father has to take care of)	아 (Oh…)	Empathic Response	Empathic Response	Continuer

Figure A.1: The samples of the backchannel prediction in the BACKSpeech dataset, ground truth backchannel category, and the prediction of CAMA-BC and BPM-V.

Figure A.2: Gradient visualization for successful sample in CAMA-BC.

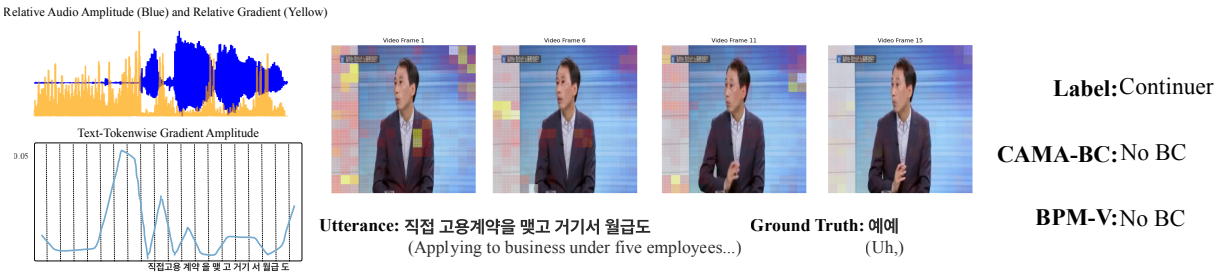


Figure A.3: Gradient visualization for failure sample in CAMA-BC.

Dataset	NoBC	BC	Train	Validation	Test	Total	#Samples for MMA-CA
SWBD	48,087	48,087	76,938	4,808	14,428	96,174	194,178

Table A.7: Distribution of each backchannel category and the number of samples used for MMA-CA. Following prior works, we ensured that the NoBC and BC samples had equal numbers. The dataset was then split into training, validation, and test sets in an $8 : 0.5 : 1.5$ ratio.

speaker. We utilized the Flux.1 dev (Black Forest Labs, 2024) model to synthesize high-quality bust images conditioned on speaker metadata. The generation prompt is defined as: *"A bust image of a* `[age]`*-year-old* `[race]` `[gender]`, `[outfit variation]`, *facing forward, highly detailed, neutral background, photo-realistic."*

Here, `[age]` and `[gender]` are extracted directly from the SWBD metadata. At the same time `[race]` and `[outfit variation]` are randomly selected from predefined lists — [*"White"*, *"Mid-*

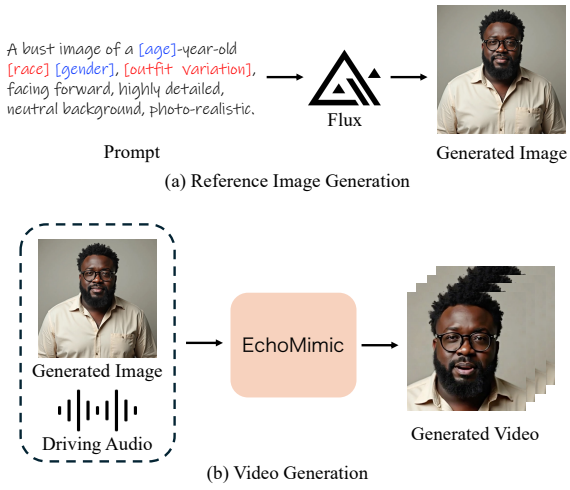


Figure A.4: Image–video generation pipeline: (a) Reference bust image generation with Flux.1 [dev] (Black Forest Labs, 2024) using diverse prompts, and (b) video generation guided by EchoMimic (Chen et al., 2025) with driving audio input.

dle Eastern", "*Black*", "*Mediterranean*"] for race and ["*wearing clothes*", "*wearing shirts*", "*wearing t-shirts*"] for outfit — to ensure visual diversity. This process is illustrated in Figure A.4 (a). An example of the generated reference images is shown in Figure A.5. Flux.1 dev achieved PDist of 0.332, SSIM of 0.896, and PSNR of 31.1, demonstrating its ability to generate high-quality, photorealistic images and supporting its suitability for our reference image generation stage.

C.2.2 Video Generation

Based on the generated reference images, we synthesize videos for the SWBD corpus to incorporate visual modality information. Aligned with the audio segments, we generate a video clip for each backchannel occurrence covering a temporal window of $n = 1500$ms following the backchannel onset. To produce realistic lip movements and facial expressions synchronized with the speech, we use Echomimic, which takes the SWBD audio as input and animates the reference image accordingly. This approach compensates for the absence of original video in SWBD and ensures consistency with the multimodal input setting used for the other datasets. The overall video generation process is illustrated in Figure A.4 (b), and examples of the generated video frames are demonstrated in Figure A.6. Echomimic achieved an FID of 29.14, FVD of 492.78, SSIM of 0.812, and E-FID of 1.11, demonstrating its ability to generate temporally consistent and visually coherent videos, which supports its suitability for our video synthesis stage.

D Additional Implementational Details

In CAMA-BC, MMA is applied at layers $\mathcal{J} = \{9, 10, 11\}$. We apply different learning rates across model components: 5×10^{-6} for encoder fine-tuning, 5×10^{-5} for classifier training, and 5×10^{-4} for randomly initialized projection and cross-attention layers. A per-epoch learning rate decay of 0.95 is applied to the encoder, projection, and cross-attention layers. All experiments were conducted using 4 NVIDIA RTX 3090 GPUs with a fixed batch size of 16. In the same computational environment, CAMA-BC requires 20.083 ms, while BPM-V requires 18.868 ms, supporting the efficiency of backchannel prediction applications. The reported inference time is measured end-to-end per sample, including feature extraction from all modalities, without feature caching.

We evaluated GPT-4.1 and Llama 4.0 Scout using carefully structured prompts that included comprehensive task descriptions and standardized response formats. To ensure systematic analysis, the models were instructed to provide explicit reasoning and confidence scores alongside their backchannel predictions, allowing for a detailed examination of their decision-making processes. The complete prompt structure for Korean datasets (KC-Dialog and BACKSpeech) is provided in D. Since SWBD uses a binary backchannel classification scheme rather than the four-class system employed in Korean datasets, we adapted the prompts accordingly, as shown in Appendix D. They showed lower performance and sensitivity to prompts for both KC-Dialog (23.71 and 26.75) and BACKSpeech (13.34 and 19.89), so we reported the Text modality as a baseline. We also evaluated a Gemini-2.5 Flash as a Multimodal LLM. For the Gemini-2.5 Flash, we utilized the same format of instructions, adding a video and audio input in `{example_transcript}` and `{request_transcript}` after the transcript; the format of multimodal input of Gemini is not publicly available.

Figure A.5: Reference image generation results with Flux.1 [dev] (Black Forest Labs, 2024) using diverse prompts for age, gender, race and outfit variations.

Generated Image Generated Video

Figure A.6: Video generation results guided by EchoMimic (Chen et al., 2025) using the driving audio and reference images.

Prompt 1 Text prompt for Llama 4.0 Scout and GPT-4.1 on KC-Dialog and BACKSpeech dataset.

```
# Korean Conversation Backchannel Analyzer
You will analyze short segments (max 5 words) of Korean conversations and
    classify the response type.
Following the instructions below.


## Definition
A backchannel is a brief response that indicates listener engagement without
     adding direct content to the conversation.

## Response Types
- **NoBC**: No backchanneling occurs
- **Continuer**: Encourages speaker to continue
- **Understanding**: Shows comprehension
- **Empathic**: Conveys emotional reaction

## Instructions
1. Analyze the given dialogue context
2. Select the most appropriate response type
3. Return your answer in JSON format:
```json
{
  "response_type": "[NoBC|Continuer|Understanding|Empathic]",
  "confidence": 0-1,
  "reasoning": "Brief explanation"
}
```

Examples

class: {example_BC_category}
Context: {example_transcript}
Response: {ground_truth}

... (n-shot for each class) ...

Now, analyze the following dialogue segment and classify the response type.
Context: {request_transcript}
```

Prompt 2 Text prompt for Llama 4.0 Scout and GPT-4.1 on SWBD dataset.

````
# Backchannel Analyzer
You will analyze short segments (max 5 words) of conversations and classify
if the response is a backchannel or not.
Following the instructions below.

## Definition
A backchannel is a brief response that indicates listener engagement
without adding direct content to the conversation.

## Response Types
- **NoBC**: No backchanneling occurs
- **BC**: Backchanneling occurs

## Instructions
1. Analyze the given dialogue context
2. Select the most appropriate response type
3. Return your answer in JSON format:
```json
{
  "response_type": "[NoBC|BC]",
  "confidence": 0-1,
  "reasoning": "Brief explanation"
}
```
## Examples

### class: {example_BC_category}
Context: {example_transcript}
Response: {ground_truth}

... (n-shot for each class) ...

Now, analyze the following dialogue segment and classify the response type.
Context: {request_transcript}
````