Breaking the Boundaries: A Unified Framework for Chinese Named Entity Recognition Across Text and Speech

Anonymous ACL submission

Abstract

In recent years, with the vast and rapidly increasing amounts of spoken and textual data, Named Entity Recognition (NER) tasks have evolved into three distinct categories, i.e., text-based NER (TNER), Speech NER (SNER) and Multimodal NER (MNER). However, existing approaches typically require designing separate models for each task, overlooking the potential connections between tasks and limiting the versatility of NER methods. To mitigate these limitations, we introduce a new task named Integrated Multimodal NER (IMNER) to break the boundaries between different modal NER tasks, enabling a unified implementation of them. To achieve this, we first design a unified data format for inputs from different modalities. Then, leveraging the pre-trained MM-Speech model as the backbone, we propose an **I**ntegrated **M**ultimod**a**l **Ge**neration Framework (**IMAGE**), formulating the Chinese IMNER task as an entity-aware text generation task. Experimental results demonstrate the feasibility of our proposed IMAGE framework in the IMNER task. Our work in integrated multimodal learning in advancing the performance of NER may set up a new direction for future research in the field.

1 Introduction

Named Entity Recognition (NER) (Li et al., 2020a) is a fundamental and significant task in the field of Natural Language Processing and has been extensively studied to address the challenges posed by real-world text data. Chinese NER (CNER) (Liu et al., 2022), a significant subdomain of NER, specifically deals with challenges unique to Chinese, such as no clear word boundaries and the ambiguity from homophones and polyphones, drawing significant academic focus (Zhang and Yang, 2018; Li et al., 2020b; Ma et al., 2020).

Traditionally, Named Entity Recognition (NER) tasks have concentrated on text-based NER (TNER) (Gui et al., 2019b; Li et al., 2020b, 2022). However, as the volume of audio data increases, there has been a growing interest in Speech NER (SNER) (Yadav et al., 2020; Chen et al., 2022; Shon et al., 2022), which focuses on extracting named entities from speech, and Multimodal NER (MNER) (Sui et al., 2021; Liu et al., 2023), which involves extracting entities from both speech and text.

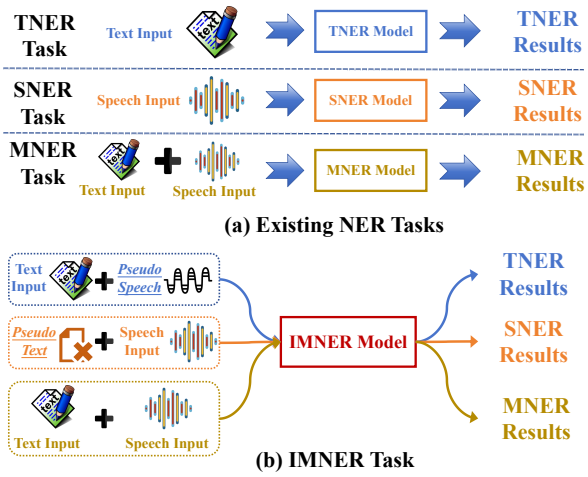


Figure 1: Comparison of existing NER tasks with the Integrated Multimodal NER (IMNER) task proposed in this paper. In the figure, TNER represents text-based NER, SNER stands for Speech NER, MNER denotes multimodal NER, Pseudo Speech refers to meaningless zero audio waveforms, and Pseudo Text indicates nonsensical text sequences.

Currently, data on the internet often appears in multiple modalities, such as user-generated content in social media and news reports in the media, which may be in text or audio formats, or a combination of both speech and its corresponding text. The key information extracted by NER tasks, such as persons and locations, can assist in search and recommendation in the news domain, as well as in analyzing trending topics and public opinion in social media. However, existing NER systems are usually designed for a single mode, either solely

as SNER, TNER or MNER, as shown in Figure 1. These approaches face two significant issues. **Issue 1**: Treating SNER, TNER and MNER as three separate tasks overlooks the potential interconnections between them. **Issue 2**: The need to design distinct models for each of the SNER, TNER and MNER tasks limits the versatility and overall efficiency of NER methods.

Beyond the above issues, significant advances in speech processing have been achieved with Multimodal Pre-trained Models (MPMs) using data from both text and speech (Zhou et al., 2022; Ao et al., 2022). These models, trained on various tasks like speech-to-text and text-to-text generation, highlight the potential interconnections between different modal tasks. However, these MPMs face a challenge, identified here as **Issue 3**: While MPMs benefit from a unified pre-training architecture across modalities, the need for task-specific fine-tuning for various downstream applications, to some extent,restricts their universality.

To address these challenges, in this paper, we introduce a new NER task named **I**ntegrated **M**ultimodal **NER** (**IMNER**). The IMNER task aims to break the boundaries of traditional SNER, TNER and MNER by presenting a unified Named Entity Recognition (NER) framework capable of handling inputs from various modalities (text, speech, or both) to efficiently recognize Chinese named entities, as illustrated in Figure 1. Moreover, previous studies (Sui et al., 2021; Liu et al., 2023) have shown that features like pauses in speech signals can reduce ambiguities in Chinese NER tasks, which often arise from the lack of clear word delimitation or the presence of homophones. The IMNER approach, leveraging data from the SNER, TNER, and MNER tasks, possesses the potential to overcome the difficulties associated with the absence of natural word segmentation and the frequent occurrence of homophones in Chinese text.

To solve the IMNER task, our approach begins with an original design of a data format unification method that transforms the data formats of TNER, SNER and MNER tasks into a unified data scheme. As illustrated in Figure 1, we treat TNER and SNER tasks as MNER tasks with missing speech and text modalities, respectively. For these "missing" modalities, we substitute Pseudo Speech and Pseudo Text. Based on the unified data format, and using the multimodal pretrained model MMSpeech (Zhou et al., 2022) as backbone, we introduce the **I**ntegrated **M**ultimod**a**l **Ge**neration Framework (**IMAGE**), an encoder-decoder structure to execute the Chinese IMNER task. Specifically, inspired by the recent success of generative methods in NER tasks, we formulate the IMNER task as an entity-aware text generation task (Chen et al., 2022; Wang et al., 2023). Unlike previous works, our approach uniquely leverages the interrelations among the three different modalities of TNER, SNER and MNER tasks, facilitating the realization of the IMNER task.

The main contributions of this work can be summarized as follows:

- We introduce a new task, Integrated Multimodal NER (IMNER), aimed at breaking the boundaries between TNER, SNER and MNER tasks, enabling the model to uniformly handle inputs from various modalities.
- From a novel perspective, we design a unified data format for TNER, SNER and MNER, establishing a bridge between these three tasks and serving as the basis for IMNER.
- Utilizing the MMSpeech model as the backbone, we propose an Integrated Multimodal Generation Framework (IMAGE), formulating the Chinese IMNER task as an entity-aware text generation task. Notably, our IMAGE framework is capable of handling both flat and nested entity scenarios.
- Experimental results reveal that the IMAGE framework effectively exploits potential connections among TNER, SNER and MNER tasks, boosting their performance. IMAGE achieves competitive performance in these tasks, proving the viability of the IMNER task and incidenting the advantages of the IMAGE framework.

2 Related Work

2.1 Text-based Chinese NER (TNER)

In Chinese NER, the lack of natural word boundaries and the existence of homophones introduce ambiguity in the text, posing challenges for Chinese NER. Therefore, in recent years, incorporating external lexicon resources to enhance Chinese NER performance has been proven to be an effective solution and has achieved significant success (Zhang and Yang, 2018; Gui et al., 2019a,b; Li et al., 2020b; Liu et al., 2021). Additionally, for the extraction of nested entities, recent work utilizing a

unified NER framework (Li et al., 2022; Yan et al., 2021) to extract both flat and nested entities has shown promising results.

2.2 Speech NER (SNER)

Speech NER (SNER), which is essential for Spoken Language Understanding (SLU) (Caubrière et al., 2020; Shon et al., 2022), initially adopts a two-stage pipeline approach (Cohn et al., 2019): converting speech to text with Automatic Speech Recognition (ASR) and then tagging named entities in the generated text. To overcome the error accumulation inherent in this approach, End-to-End (E2E) methods for languages like French (Ghannay et al., 2018), English (Yadav et al., 2020), and Chinese (Chen et al., 2022) have emerged, which incorporate entity-aware ASR, directly integrating entity tagging into the ASR decoding process.

2.3 Multimodal NER (MNER)

With the rapid growth of multimodal data on the internet, leveraging multimodal information to enhance the performance of NER systems has attracted increasing academic attention. In the field of English NER, existing work has primarily focused on using data from text and image modalities to improve the performance of NER systems in social media contexts (Sun et al., 2021; Chen et al., 2021; Xu et al., 2022; Jia et al., 2023). Similarly, in the field of Chinese NER, Multimodal NER (MNER) that combines text with audio signals (Sui et al., 2021; Liu et al., 2023) has been introduced and achieved significant success.

2.4 MPMs Based on Text and Speech

Recently, Multimodal Pretrained Models (MPMs) have received widespread attention in the field of speech processing. In English, models such as SpeechT5 (Ao et al., 2022) and STPT (Tang et al., 2022), which propose encoder-decoder pre-training using unlabeled text and speech data, have achieved significant success. Following this, in the Chinese Automatic Speech Recognition (ASR), MMSpeech (Zhou et al., 2022) makes great improvements through a multi-modal multi-task encoder-decoder pre-training framework.

It is important to note that, in our work, although MMSpeech serves as the backbone, our model, named IMAGE, distinguishes itself from the aforementioned MPMs by overcoming the need for individual fine-tuning across different downstream tasks. Furthermore, while recent works like SpeechGPT (Zhang et al., 2023) and Qwen-Audio (Chu et al., 2023) have demonstrated the capability to handle both speech and text inputs in conversational tasks, to our knowledge, our work is the first attempt to explore integrated modality input capability within an NER system.

3 Methodology

In this section, we first introduce the IMNER task definition. Then we detail the implementation of IMAGE, including its overal structure as depicted in Figure 2.

3.1 Task Definition

Given the input X, which may be text $\{\mathbf{x}_{\text{text}}\}$, speech $\{\mathbf{x}_{\text{speech}}\}$, or a combination of both $\{\mathbf{x}_{\text{text}}, \mathbf{x}_{\text{speech}}\}$, the goal of IMNER is to find each entity in X and then assign a label $y \in Y$, where Y is a predefined label types (e.g., PER, LOC, etc.).

3.2 Formulating IMNER into Text Generation

Inspired by the success of generative methods in TNER (Wang et al., 2023) and SNER (Chen et al., 2022), we formulating the IMNER task as an entity-aware text generation task. Illustrated by the sample in Figure 2, for the text "末阳市文联主席张三" and its associated speech waveform, the Entity-aware Text Generation Target is designated as "<(末阳市)文联>主席[张三]". Special tokens are incorporated into the vocabulary to annotate entities in the generated text, specifically, "[ ]" for PER, "( )" for LOC, and "< >" for ORG. We chose the entity-aware text generation task for generating entities primarily because this method allows for the simultaneous acquisition of entity span information and entity text content.

3.3 Details of the IMAGE Framework

3.3.1 Backbone Model

In this paper, we employ a multimodal pre-trained model with an encoder-decoder structure, MMSpeech (Zhou et al., 2022), as our backbone model. The original MMSpeech structure primarily consists of: (1) a multi-layer Transformer-based MMSpeech encoder shared by text and speech modalities, equipped with a multi-layer convolutional and Transformer-based speech feature extractor, and a static word vector embedding for text feature extraction; (2) a decoder composed of multiple Transformer layers.

Currently, MMSpeech is mainly used for downstream tasks with speech input, such as speech

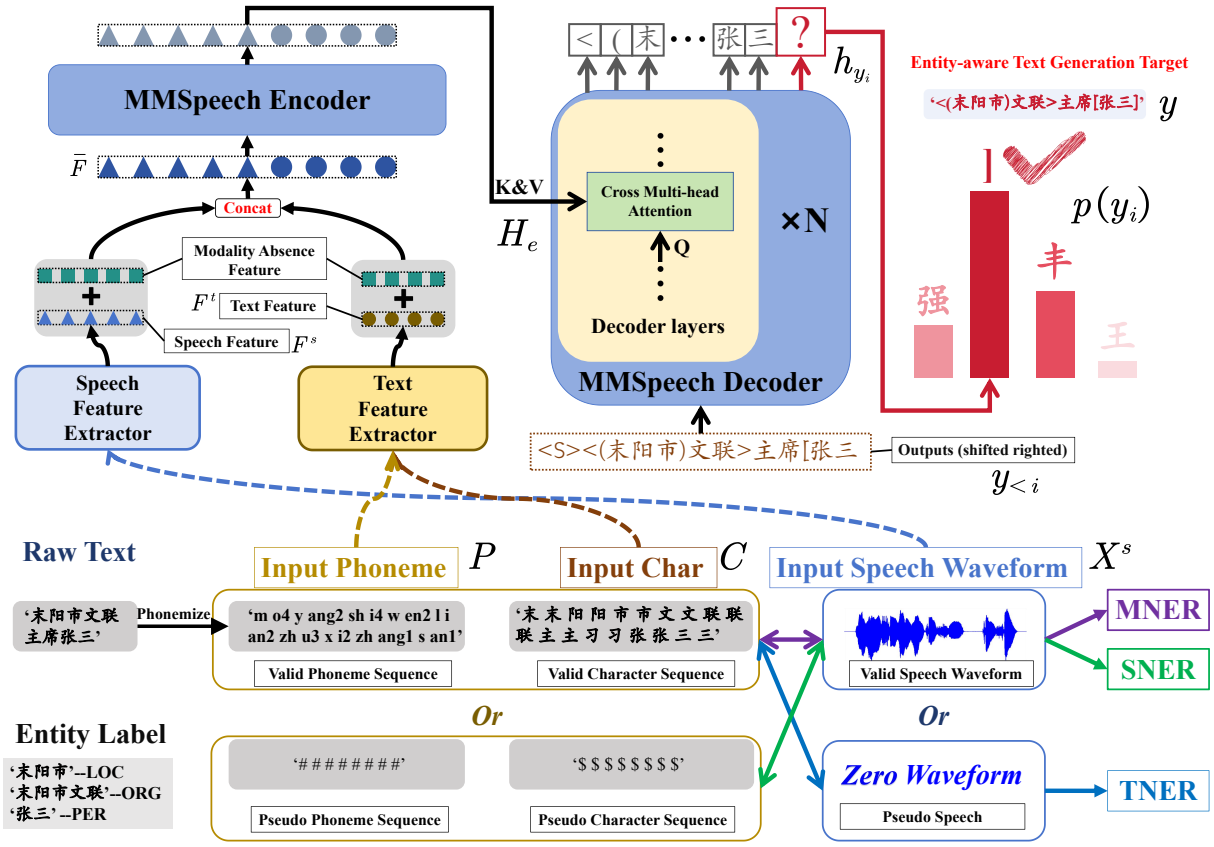


Figure 2: Overall structure of IMAGE. The figure illustrates with the example of the text "末阳市文联主席张三"(Chairman Zhang San of the Moyang City Cultural Association) and its corresponding speech waveform. In the figure, "<S>" denotes a special token indicating the start of the generated output, while "#" and "$" respectively represent meaningless phoneme tokens and Chinese character tokens. The purple, green, and blue arrows at the bottom right of the figure explain the composition of input data for the MNER, SNER, and TNER tasks within the IMAGE framework, respectively.

recognition. Besides speech data, MMSpeech was trained with a large volume (292GB) of text data, giving it a strong ability to model text. However, the capability of MMSpeech to handle text input tasks or dual input tasks with speech and text has been overlooked and not fully explored. This indicates the significant potential for expanding MMSpeech's application across various modal scenarios.

3.3.2 Unified Integrated Modal Data Format

The IMNER task comprises three sub-tasks: SNER, TNER and MNER, each involving different modal components in the input data. To transform the integrated modal inputs of IMNER task into a unified data format, maintaining data consistency and laying the groundwork for handling all three tasks with a uniform model structure, we adopt a novel perspective within the IMAGE framework. Here, we treat TNER and SNER tasks as MNER tasks with "missing" speech and text modalities, respectively. For these "missing" modalities, we substitute Pseudo Speech and Pseudo Text as illustrated in Figure 2.

In the Unified Integrated Modal Data Format, the Input Speech Waveform is denoted as $X^s = \{x_1^s, \cdots, x_{N^s}^s\}$, where N^s represents the length of the speech waveform. When the speech modality is missing from the input, X^s represents a fixed-length sequence of all-zero signals. Because Chinese characters and their corresponding sounds are not tightly mapped to one another, the encoder in MMSpeech converts the original text input $X^t = \{x_1^t, \cdots, x_{N^t}^t\}$ into phoneme input $P = \{p_1, \cdots, p_{N^p}\}$, where N^t and N^p denote the sequence lengths of X^t and P, respectively.

However, as mentioned in Section 3.2, our entity generation scheme requires to restore all raw input text while generating special tokens for annotated entities in decoder. Through practice, we observed that with lots of homophones in Chinese, the decoder of original MMSpeech might not always ac-

curately restore the input text solely based on the input phoneme sequence sometimes. For example, the MMSpeech Decoder might restore the input "末阳市(Mo Yang Shi)" as "末杨市(Mo Yang Shi)" or "西安站(Xi An Zhan)" as "鲜站(Xian Zhan)", leading to incorrect entity recognition. To address this issue, we introduce a sequence of Chinese characters $C = \{c_1, \cdots, c_{N^p}\}$ (where c_i is the Chinese character corresponding to the phoneme p_i) as auxiliary features to assist the model in accurately restoring the input text at the output end of MMSpeech Decoder. Specifically, for Pseudo Text, we choose to represent P and C using semantically meaningless strings, such as "########" for P and "$$$$$$$$" for C, as shown in Figure 2.

3.3.3 Feature Extractors

Speech Feature Extractor: Consistent with MMSpeech, we first convert the raw speech waveform into Mel-filterbank features, and then use a multi-layer convolutional network followed by a Transformer encoder (comprising multiple transformer layers with multihead self-attention (Vaswani et al., 2017)) as the speech feature extractor. The speech features F^s for the input speech X^s are computed as follows:

$$F^s = SFE(X^s) \tag{1}$$

where $SFE(\cdot)$ denotes the Speech Feature Extractor, $F^s \in \mathbb{R}^{L_{F^s} \times d_h}$, L_{F^s} is the length of the speech features, and d_h is the dimension of the hidden features (consistent with all d_h in this paper).

Text Feature Extractor: Consistent with the pre-training phase of MMSpeech, we utilize static embeddings in our model to obtain the feature representation of the input phoneme sequence P:

$$F^p = E^{(p)}(P) \tag{2}$$

where $E^{(p)}(\cdot)$ denotes the operation of static phoneme embedding, and $F^p \in \mathbb{R}^{N^p \times d_h}$.

Differing from the original MMSpeech, as described in Section 3.3.2, we introduce a character feature sequence C as an auxiliary feature to enhance the model's ability to perceive Chinese characters and improve its capability to accurately restore input Chinese characters at the decoding output. We also use static embeddings to obtain the feature representation of C:

$$F^c = E^{(c)}(C) \tag{3}$$

where $E^{(c)}(\cdot)$ denotes the operation of static character embedding, and $F^c \in \mathbb{R}^{N^p \times d_h}$.

We then add F^p and F^c together to derive the final text feature F^t:

$$F^t = F^c + F^p \tag{4}$$

Modality Absence Feature: To enhance the model's ability to detect the absence of a modality, thereby encouraging it to focus on inputs from present modalities and ignore inputs from missing ones, we introduce a learned embedding to every $f_i^t (1 \leqslant i \leqslant N^p)$ and $f_i^s (1 \leqslant i \leqslant N^s)$ to incident whether that modality is missing:

$$\bar{f}_i^t = \begin{cases} f_i^t + m_{\text{missing}}, & \text{if text is missing} \\ f_i^t + m_{\text{present}}, & \text{otherwise} \end{cases} \tag{5}$$

$$\bar{f}_i^s = \begin{cases} f_i^s + m_{\text{missing}}, & \text{if speech is missing} \\ f_i^s + m_{\text{present}}, & \text{otherwise} \end{cases} \tag{6}$$

where $m_{\text{missing}} \in \mathbb{R}^{d_h}$ and $m_{\text{present}} \in \mathbb{R}^{d_h}$ are the learned embeddings indicating the absence or presence of the modality, respectively. This approach allows the model to dynamically adapt its processing based on the availability of each modality. Ultimately, we obtain the final speech feature representation $\bar{F}^s = \left\{\bar{f}_1^s, \cdots, \bar{f}_{N^s}^s\right\}$ and the final text feature representation $\bar{F}^t = \left\{\bar{f}_1^t, \cdots, \bar{f}_{N^p}^t\right\}$.

3.3.4 Encoder and Decoder of MMSpeech

For $\bar{F}^s$ and $\bar{F}^t$, we combine them through a concatenation operation to form the feature representation $\bar{F}$ that is fed into the MMSpeech encoder-decoder structure:

$$\bar{F} = \bar{F}^t \oplus \bar{F}^s \tag{7}$$

where $\bar{F} \in \mathbb{R}^{(N^p+N^s) \times d_h}$, and $\oplus$ denotes the concatenation operation.

Encoder: IMAGE feeds the concatenated text and speech features representation $\bar{F}$ into the MMSpeech encoder, which is a multi-layer Transformer encoder, to obtain the hidden representation of the integrated modal input as follows:

$$H_e = Encoder\left(\bar{F}\right) \tag{8}$$

Decoder: Afterwards, H_e is fed into the MMSpeech decoder, a multi-layer Transformer decoder, to model the probability distribution of the output text y. At the i-th step of decoding, the probability distribution $p\left(y_i\right) \in \mathbb{R}^{|V|}$ of the i-th output token y_i in y is computed as follows:

$$h_{y_i} = Decoder\left(H_e, y_{<i}\right) \tag{9}$$

$$p\left(y_i\right) = Softmax\left(W_{lm} h_{y_i} + b_{lm}\right) \tag{10}$$

Dataset	Data Type	Entity Type Num	Sentence				Entity Type			
			train	dev	test	total	PER.	ORG.	LOC.	total
CNERTA	Text&Audio	3	34,102	4,440	4,445	42,987	8,034	12,047	16,876	36,957
AISHELL-NER	Text&Audio	3	120,098	14,326	7,176	141,600	18,642	25,351	24,611	68,604
MSRA	Text-only	3	46,539	-	4,380	73,321	18,565	21,804	32,952	73,321

Table 1: Statistics of the Datasets.

where h_{y_i} is the hidden representation at the i-th decoding step, $W_{lm} \in \mathbb{R}^{|V| \times d_h}$ and $b_{lm} \in \mathbb{R}^{|V|}$ are learnable parameters in the language model (LM) head, and $|V|$ represents the size of the vocabulary.

3.3.5 Training Strategy of IMAGE

Loss Function: During the training phase, the parameters of IMAGE are optimized by minimizing the cross-entropy loss based on teacher forcing:

$$\mathcal{L} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{|V|} l_{i,k} \log p\left(y_{i,k}\right) \qquad (11)$$

where $l_i \in \mathbb{R}^{|V|}$ represents the ground-truth label distribution for decoding the i-th token, and M denotes the total number of tokens in the ground-truth label sentence.

Training Data Creation: In this study, we utilize MNER datasets, containing both speech and text, to create training data for TNER and SNER by artificially removing certain modality data. During training, each sample in a batch is randomly assigned as an input for MNER, SNER or TNER. These input variations are depicted in Figure 2, illustrating the method for handling inputs with varying modality presence.

4 Experiments

4.1 Dataset & Evaluation Metrics

In this study, we train and evaluate our IMAGE framework on the IMNER task using three datasets: the flat SNER dataset AISHELL-NER (Chen et al., 2022) , the nested MNER dataset CNERTA (Sui et al., 2021), and the flat TNER dataset MSRA (Levow, 2006). Both the AISHELL-NER and CNERTA datasets contain Chinese text with corresponding speech, where the Chinese text is annotated with entity information, while the MSRA dataset solely contains Chinese text with entity annotations. Detailed statistics of these datasets are available in Table 1 in in Appendix. Regarding evaluation metrics, we use the F1 score (F1), commonly employed in NER tasks, to assess the model's effectiveness. **For further implementation details, see Appendix Section A .**

4.2 Comparison Models

We compare the performance of several strong baseline models on the TNER, SNER and MNER tasks using the benchmark datasets employed in this paper. The baseline models used fall into three main categories: (1) methods that only use the text modality (Text-only Methods), (2) methods that only use the speech modality (Speech-only Methods), and (3) multimodal methods that use both text and speech (Multimodal Methods).

Due to page limitations, detailed introductions of each baseline model can be found in Appendix Section B .

4.3 Results and Analysis

4.3.1 Main Results

We compare our proposed IMAGE model with several strong Text-only Baselines, Speech-only Baselines, and Multimodal Baselines, with the experimental results reported in Table 2. It is evident that, unlike existing baseline models that are limited to solving single-modality tasks, IMAGE not only breaks the boundaries between modalities by simultaneously addressing TNER, SNER and MNER tasks but also achieves highly competitive performance across these tasks. From the experimental results, we can further observe that:

(1) Compared to baselines using MMSpeech trained on single task data (methods 6, 7, 9, 10, 16, 17 in Table 2), our proposed IMAGE method achieves significant performance improvements on TNER, SNER and MNER tasks. This demonstrates that our IMAGE method can effectively leverage the potential correlations among the three tasks across different input modalities, facilitating complementary benefits and jointly enhancing performance across all tasks.

(2) On the SNER and MNER tasks, the performance of IMAGE with the MMSpeech-large backbone surpasses all baseline methods. Furthermore, IMAGE with the MMSpeech-base backbone, despite having only about 213M parameters, still remains competitive. This indicates that our proposed IMAGE framework can exploit the comple-

Modality	Methods	CNERTA(nested)			AISHELL-NER		
		TNER	SNER	MNER	TNER	SNER	MNER
Text-only Methods	1.Bert-large-CRF(325M)	76.09$^{\hbar}$	-	-	93.29$^{\hbar}$	-	-
	2.FLAT(Bert-large)	79.31$^{\ddagger\hbar}$	-	-	93.57$^{\ddagger\hbar}$	-	-
	3.W^{2}NER(Bert-large)	79.25‡	-	-	**93.72**‡	-	-
	4.Bart-large-ETG(407M)	76.84	-	-	92.82	-	-
	5.MT5-base-ETG(582M)	76.91	-	-	92.94	-	-
	6.MMSpeech-base-ETG(213M)	75.93	-	-	90.23	-	-
	7.MMSpeech-large-ETG(613M)	76.90	-	-	92.81	-	-
Speech-only Methods	8.Conformer-ETG(E2E)(Chen et al., 2022)	-	60.36‡	-	-	73.37$^{\wp}$	-
	9.MMSpeech-base-ETG(E2E)	-	67.21	-	-	74.28	-
	10.MMSpeech-large-ETG(E2E)	-	70.35	-	-	75.42	-
	11.Conformer-ASR + Bert-large(Pipeline)	-	60.92	-	-	74.10	-
	12.MMSpeech-base-ASR + Bert-large(Pipeline)	-	66.87	-	-	73.14	-
	13.MMSpeech-large-ASR + Bert-large(Pipeline)	-	69.76	-	-	74.84	-
Multimodal Methods	14.Bert-USAF(Bert-base)(Liu et al., 2023)	-	-	76.73$^{\wp}$	-	-	-
	15.Bert-M3T(Bert-large)	-	-	79.51$^{\ddagger\hbar}$	-	-	93.75$^{\ddagger\hbar}$
	16.IMAGE(MMSpeech-base,only MNER data)	-	-	76.68	-	-	90.84
	17.IMAGE(MMSpeech-large,only MNER data)	-	-	80.79	-	-	93.69
IMNER Methods	18.IMAGE(MMSpeech-base,213M)	76.61	68.27	76.96	90.95	75.35	91.10
	19.IMAGE(MMSpeech-large,613M)	**80.83**	**71.17**	**81.10**	93.49	**76.33**	**93.83**

Table 2: F1-score (%) of the proposed IMAGE method and baselines on the TNER, SNER, MNER versions of the test sets for two benchmark datasets. Here, "ETG" refers to models perform NER task using an entity-aware text generation task. "(E2E)" and "(Pipeline)" respectively denote the end-to-end SNER methods and pipeline SNER methods. Superscript ‡ indicates results obtained through official implementation. Superscript $\wp$ denotes experimental results reported from the original paper. Superscript $\hbar$ signifies the use of the same Nested Structure Linearization method for annotating nested entities as in the M3T(Sui et al., 2021) work.

mentarity between tasks across different modalities, enhancing modeling capabilities for both speech-only and speech-text multimodal tasks.

(3) On the TNER task, our IMAGE method with MMSpeech-large backbone achieves performance comparable to the SOTA method W^{2}NER. Moreover, the performance of the IMAGE framework exceeds all baselines based on an encoder-decoder structure using an entity-aware text generation task for entity annotation. This suggests that the IMAGE framework can improve text information encoding capabilities through joint training and modeling of tasks across different input modalities.

(4) The performance of the IMAGE method on the MNER task surpasses its performance on the TNER task, indicating that multimodal inputs combining text and speech provide more effective information than text-only data, thus enhancing model performance on NER tasks. Additionally, the SNER task not only requires entity annotation but also the accurate transcription of speech to text, increasing the complexity of the task. Therefore, the performance of the IMAGE framework on the SNER task is notably lower than on the TNER task.

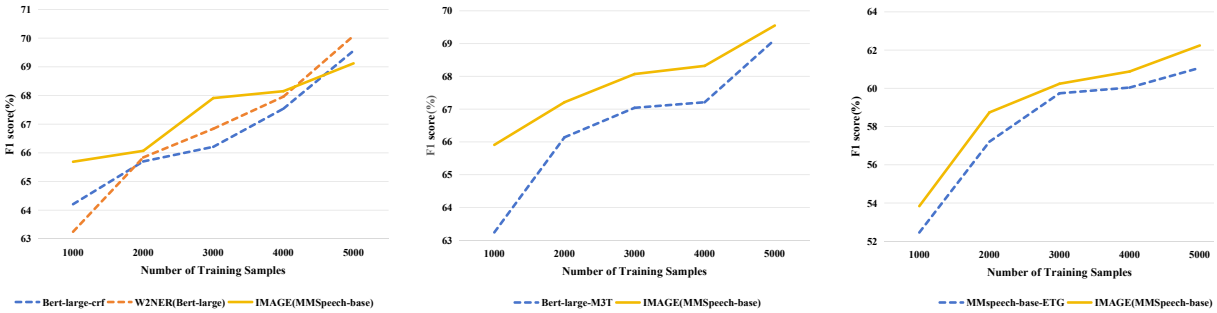
4.3.2 Ablation Study

To assess the impact of various components in IMAGE, we conducted ablation experiments on the CNERTA dataset, with the findings presented in Table 3. Our conclusions are as follows:

(1) Removing the Character Feature F^c from IMAGE results in a performance decline, particularly in the TNER and MNER tasks. This indicates that the original MMSpeech model, which relies solely on phoneme input for the text modality, might introduce errors in the decoding of Chinese characters. Incorporating the Character Feature improves the model's ability to interpret the input Chinese characters, thus alleviating this problem.

(2) The removal of the Modality Absence Feature leads to reduced model performance, highlighting its role in enhancing IMAGE's ability to discern valid modal information in the input. Additionally, eliminating either Pseudo Text Input or Pseudo Speech Input diminishes the model's performance. These components are believed to capture global information across different modalities, fostering synergy among the tasks within IMAGE.

(3) Excluding the training data for any one of the TNER, SNER, and MNER tasks during the training process results in a decline in the model's performance across all three NER subtasks. This suggests



(a) Results on the TNER Task **(b) Results on the MNER Task** **(c) Results on the SNER Task**

Figure 3: F1 Scores (%) of IMAGE (MMSpeech-base) on the three subtasks of the IMNER task in the CNERTA dataset with varying numbers of training samples. For each method depicted above, under different training data sizes, we chose identical hyperparameters and ran the experiments five times with different random seeds, averaging the F1 scores to obtain the final results.

	TNER	SNER	MNER
IMAGE	80.83	71.17	81.10
w/o CF	78.51	71.10	79.04
w/o MAF	80.12	70.61	80.57
w/o PT	80.62	70.69	80.98
w/o PS	80.25	71.02	80.89
w/o PT&PS	80.21	70.59	80.81
w/o TNER	40.35	70.35	80.86
w/o SNER	80.67	36.27	80.91
w/o MNER	80.59	70.21	42.21

Table 3: An ablation study of the IMAGE (MMSpeech-large). F1 scores (%) were evaluated on the test sets of three different tasks in CNERTA. "CF" stands for Character Feature F^c. "MAF" represents Modality Absence Feature. "PT" and "PS" respectively denote Pseudo Text Input and Pseudo Speech Input. The feature vectors corresponding to "PT" and "PS" are masked out in the Transformer through the attention mask to negate their influence. The acronyms "TNER", "SNER", and "MNER" specifically refer to the training data for the respective tasks.

that there is an intrinsic interconnection among the TNER, SNER, and MNER tasks, allowing them to mutually benefit from each other. This finding supports our IMNER approach, addressing **Issue 1** raised in the introduction of this paper, which concerns the overlooked potential interconnections between these tasks.

4.3.3 Performance on Low-resource Scenarios

We conducted experiments with training data sets of 1000, 2000, 3000, 4000, and 5000 samples to evaluate the performance of IMAGE in low-resource scenarios, with results shown in Figure 3. The experimental results reveal that, with limited training data resources, IMAGE maintains an advantage compared to baselines trained on single-task data. This demonstrates that within the low-resource context, the IMAGE framework can still effectively leverage the potential connections and complementarity among the three IMNER subtasks (i.e., TNER, SNER, MNER) to enhance the performance across these tasks. Notably, on the TNER task, when the training data ranges from 1000 to 4000 samples, the performance of the IMAGE method using MMSpeech-base (213M) surpasses that of the baseline method using a larger backbone model, Bert-large (325M). This underscores the potential of the IMAGE method in scenarios with limited training resources.

Due to page limitations, additional experimental results and analysis have been included in the Appendix Section.

5 Conclusions

In our study, we introduce the Integrated Multimodal NER (IMNER) task, bridging the gap between text-based NER, speech NER, and multimodal NER to enable a unified approach to these three distinct tasks. By designing a novel unified data format and leveraging the pre-trained MMSpeech as backbone, we introduced the IMAGE framework, transforming the Chinese IMNER task into an entity-aware text generation task. Experimental results reveal the effectiveness of IMAGE, marking a significant step forward in integrated multimodal learning for NER, which may shed light on future research in this research domain.

Limitations

In this section, we discuss two limitations of the IMAGE framework as follows:

(1) Language Limitation: Currently, the IMAGE framework is designed to address the Chinese IMNER task exclusively. This restriction arises because the MMSpeech backbone, on which IMAGE relies, exhibits robust and balanced representation capabilities in both text and speech modalities only in Chinese. In contrast, English lacks multimodal pre-training models that perform equally well across both modalities. The available models, such as SpeechT5 (Ao et al., 2022) and STPT (Tang et al., 2022), have been pre-trained on limited text corpora, resulting in weaker text representation capabilities. Therefore, there is an urgent need to develop multimodal pre-trained models using extensive text and speech data in other languages, such as English, to support IMNER tasks in those languages.

(2) Task Limitation: At present, the IMAGE framework has only been applied to the Chinese Integrated Multimodal Named Entity Recognition (IMNER) task. Future work will involve extending the IMAGE framework to other integrated multimodal information extraction tasks. This expansion aims to fully exploit the complementary nature of different modality tasks, enhancing the overall performance and applicability of the framework.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738.

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in named entity recognition from speech? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4514–4520.

Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang. 2022. Aishell-ner: Named entity recognition from chinese speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8352–8356. IEEE.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Can images help recognize entities? a study of the role of images for multimodal ner. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 87–96.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran, Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2019. Audio de-identification: A new entity recognition task. In *Proceedings of NAACL-HLT*, pages 197–204.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019a. Cnn-based chinese ner with lexicon rethinking. In *IJCAI*.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuan-Jing Huang. 2019b. A lexicon-based graph neural network for chinese ner. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1040–1050.

Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. 2023. Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8032–8040.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN workshop on Chinese language processing*, pages 108–117.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020a. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022.

Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020b. Flat: Chinese ner using flat-lattice transformer. In *ACL*.

Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li. 2022. Chinese named entity recognition: The state of the art. *Neurocomputing*, 473:37–53.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced chinese sequence labeling using BERT adapter. In *ACL*, pages 5847–5858.

Ye Liu, Shaobin Huang, Rongsheng Li, Naiyu Yan, and Zhijuan Du. 2023. Usaf: Multimodal chinese named entity recognition using synthesized acoustic features. *Information Processing & Management*, 60(3):103290.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuan-Jing Huang. 2020. Simplify the usage of lexicon in chinese ner. In *ACL*, pages 5951–5960.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, and Jun Zhao. 2021. A large-scale chinese multimodal ner dataset with speech clues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2807–2818.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, et al. 2022. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1488–1499.

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.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: a general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1215–1223.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4268–4272. ISCA.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various ner subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822.

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.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *ACL*, pages 1554–1564.

Xiaohuan Zhou, Jiaming Wang, Zeyu Cui, Shiliang Zhang, Zhijie Yan, Jingren Zhou, and Chang Zhou. 2022. Mmspeech: Multi-modal multi-task encoder-decoder pre-training for speech recognition. *arXiv preprint arXiv:2212.00500*.

A Experiment Settings

During the training phase, we generate each training sample according to the Training Data Creation method described in Section 3.3.5. In the evaluation phase, we first manually remove the corresponding modality information from the test set to produce three versions of the test set for TNER,

SNER and MNER, allowing for a comprehensive evaluation of the model's performance.

Additionally, for the MSRA dataset, which follows the same annotation guidelines as the AISHELL-NER dataset and includes the same entity types, we employ the entire MSRA training set (46,539 samples) along with all the SNER training data from the AISHELL-NER dataset (120,098 samples) for training our model. Subsequently, we evaluate the model's performance on the TNER task using MSRA's test set and on the SNER task using AISHELL-NER's test set. The rationale behind this experimental setup is to verify the effectiveness of the unified cross-modal training strategy in IMAGE on the widely used TNER dataset, i.e., MSRA.

For our IMAGE model, we initialize the model parameters using the pre-trained MMSpeech model, including components such as the Speech Feature Extractor, Text Feature Extractor, MMSpeech Encoder, and MMSpeech Decoder. We trained IMAGE using both the Base[1] and Large[2] versions of the MMSpeech pre-trained model weights, employing teacher forcing and reported results for both. During the training phase, the training data for MNER, SNER, and TNER were balanced with a ratio of 1:1:1. For the Base version, we used a batch size of 24 and a learning rate of 3e-5. For the Large version, we used a batch size of 12 and a learning rate of 1e-5. During the decoding phase, we applied the beam search method with a beam width of 5. Additionally, we incorporated Pseudo Phoneme and character equences of length 40 in our model. For the pseudo speech input, we utilized a sequence of zero values with a length of 10,000. We implemented the training of the IMAGE model using PyTorch on a GeForce RTX 4090 GPU, employing the AdamW optimizer with a warm-up rate of 0.1, and trained it for 50 epochs on each dataset.

B Comparison Models

We compare the performance of several strong baseline models on the TNER, SNER and MNER tasks using the benchmark datasets employed in this paper. The baseline models used fall into three main categories: (1) methods that only use the text modality (Text-only Methods), (2) methods that only use the speech modality (Speech-only Methods), and (3) multimodal methods that use both text and speech (Multimodal Methods). The introductions to the three categories of baseline models selected for our comparison are as follows:

(1) **Text-only Methods**: We chose two types of baselines in this category. The first type includes State-of-the-Art (SOTA) methods based on Bert-large[3] (Cui et al., 2020), such as *Bert-large-CRF*, *FLAT* (Li et al., 2020b), and *W²NER* (Li et al., 2022). Bert-large-CRF and FLAT employ the same Nested Structure Linearization method for annotating nested entities as in the M3T(Sui et al., 2021) work. The second type consists of models with an encoder-decoder structure for NER using an entity-aware text generation task (ETG), including *Bart-large*[4] (407M) (Shao et al., 2021), *MT5-base*[5] (582M) (Xue et al., 2021), and the original version of *MMSpeech* (which uses only Chinese phoneme input at the encoder).

(2) **Speech-only Methods**: We also chose two types of baselines in this category. The first type includes end-to-end methods based on an encoder-decoder structure, where the encoder inputs are speech, and the decoder annotates entities using an entity-aware text generation task, including *Conformer-ETG* (Chen et al., 2022) and *MMSpeech-ETG*. The second type involves Pipeline methods that first recognize speech into text using ASR and then annotate entities using Bert-large, such as *Conformer-ASR + Bert-large* (Chen et al., 2022), *MMSpeech-base-ASR*[6] *+ Bert-large*, *MMSpeech-large-ASR*[7] *+ Bert-large*.

(3) **Multimodal Methods**: We selected the SOTA methods in Multimodal Named Entity Recognition (MNER) based on speech and text, including *Bert-USAF* (Sui et al., 2021) and *Bert-M3T* (Liu et al., 2023). Additionally, we included the *IMAGE framework trained solely on MNER data* as a baseline method.

C Results on Widely Used MSRA Dataset

To validate the performance of our proposed IMAGE framework on the MSRA dataset, which is

[1] https://www.modelscope.cn/models/iic/ofa_mmspeech_pretrain_base_zh

[2] https://modelscope.cn/models/iic/ofa_mmspeech_pretrain_large_zh

[3] https://huggingface.co/hfl/chinese-macbert-large

[4] https://huggingface.co/fnlp/bart-large-chinese

[5] https://huggingface.co/google/mt5-base

[6] https://modelscope.cn/models/iic/ofa_mmspeech_asr_aishell1_base_zh

[7] https://modelscope.cn/models/iic/ofa_mmspeech_asr_aishell1_large_zh

Modality	Methods	AISHELL-NER (SNER)	MSRA (TNER)
Text-only Methods	Bert-large-CRF(325M)	-	95.08
	FLAT(Bert-large)	-	**96.23**
	W2NER(Bert-large)	-	96.18
	Bart-large-ETG(407M)	-	95.21
	MT5-base-ETG(582M)	-	95.46
	MMSpeech-base-ETG(213M)	-	94.32
	MMSpeech-large-ETG(613M)	-	95.48
Speech-only Methods	MMSpeech-base-ETG(E2E)	74.28	-
	MMSpeech-large-ETG(E2E)	75.42	-
IMNER Methods	IMAGE(MMSpeech-base,213M)	74.76	94.91
	IMAGE(MMSpeech-large,613M)	**75.92**	96.14

Table 4: F1-scores (%) for the SNER task on the AISHELL-NER dataset and the TNER task on the MSRA dataset. The IMNER method is trained jointly using SNER data from the AISHELL-NER dataset and TNER data from the MSRA dataset.

	TNER	SNER	MNER
STDRC	80.83	71.17	81.10
BTDRC	80.60	71.03	80.91
STLT	79.49	70.43	80.17

Table 5: F1 Scores (%) of IMAGE (MMSpeech-large) on the CNERTA dataset using different training data creation strategies.

widely used in the NER domain, we trained the model as described in the second paragraph of Section A. The experimental results are presented in Table 4. The findings reveal that using the data from both the AISHELL-NER dataset's SNER task and the MSRA dataset's TNER task for joint training significantly improves the model's performance on both datasets. However, due to the large difference in sample length distribution between the two datasets, with the AISHELL-NER dataset tending to have shorter text lengths and the MSRA dataset featuring a more uniform length distribution, the performance improvement of the jointly trained model is not as substantial as the improvements detailed in Table 2. Additionally, the experimental results show that the IMAGE framework exhibits competitive performance on the widely used MSRA dataset.

D Analysis of Different Training Data Creation Strategies

As described in the main text of the paper, during training, each sample in a batch is randomly designated as an input for MNER, SNER or TNER. We refer to this training data creation approach as Sample-level Training Data Random Creation (STDRC). Furthermore, we conducted experiments to assess the impact of other training data generation strategies on model performance, with results shown in Table 5. Two alternative strategies were examined for comparison. The first is called Batch-level Training Data Random Creation (BTDRC), which differs from STDRC in that, during training, all samples in the same batch are randomly assigned as one type of input data for MNER, SNER or TNER. The second approach, Sequential Task-level Training (STLT), involves using training data in a sequential order of MNER, SNER or TNER during each epoch of the training process. Experimental results indicate that the performance of IMAGE using the STDRC strategy surpasses that of both BTDRC and STLT. This suggests that STDRC enables IMAGE to more effectively learn the potential connections between the MNER, SNER and TNER tasks, thereby enhancing the model's generalization capability.

E Impact of Data Ratio Variability on Model Performance

In the main text of the paper, we initially mixed the SNER, TNER, and MNER data in a 1:1:1 ratio to train the model. To assess the impact of varying data ratios across these three tasks on model performance, we designed an experiment where we fixed the training data volume for two tasks while progressively reducing the training data for the third task, and observed changes in model performance. The results of this experiment can be found in Figure 4. The findings indicate that the training data

Speech Model encoder	Speech Model decoder	Text Model	Winner
Transformer (pt,W2V2, HuBERT)	CTC (wpt)	Transformer (pt,DeBERTa)	Pipeline(Shon et al., 2022)
Transformer (wpt)	Transformer (wpt)	Transformer (pt,BERT)	Pipeline(Chen et al., 2022)
Transformer (wpt)	Transformer (wpt)	Transformer (wpt,BERT)	E2E(Chen et al., 2022)
Conformer (wpt)	Conformer (wpt)	Transformer (pt,BERT)	Pipeline(Chen et al., 2022)
Conformer (wpt)	Conformer (wpt)	Transformer (wpt,BERT)	E2E(Chen et al., 2022)
Transformer (pt,MMSpeech encoder)	Transformer (pt,MMSpeech decoder)	Transformer (pt,BERT)	E2E(ours)

Table 6: Comparison of the performance of End-to-End and Pipeline methods on the SNER task between existing work and this study. "pt" denotes finetuning with pretrained model weights, while "wpt" indicates training with randomly initialized weights.

from the SNER, TNER, and MNER tasks all contribute to enhancing model performance, and reducing the training data for any of the NER subtasks results in a performance decline in the IMAGE framework. This reinforces the motivation discussed in the introduction that joint training across the SNER, TNER, and MNER tasks can be mutually beneficial.

F End-to-End vs. Pipeline on SNER task

Previous work on SNER (Shon et al., 2022; Chen et al., 2022) compared the performance between End-to-End (E2E) SNER methods and Pipeline methods. Initially, when speech models for ASR utilize pretrained Transformer encoders (like W2V2 and HuBERT) with CTC decoders, the performance of E2E methods tends to be inferior to Pipeline methods due to the shallow layers of CTC decoders and their weak text decoding capabilities. However, when speech models for ASR (with an encoder-decoder structure) employ unpretrained full Transformer or Conformer for both encoder and decoder, the increased layers in the decoder enhance its expressive and language modeling capabilities, improving text decoding performance and thereby allowing E2E methods to surpass Pipeline methods that use randomly initialized BERT text models. Yet, if the Pipeline method employs a pretrained BERT text model, its performance exceeds that of E2E models (as shown in (Chen et al., 2022) and methods 11 and 12 in Table 2). To our knowledge, the performance comparison between End-to-End and Pipeline on the SNER task using a pretrained full Transformer encoder-decoder ASR model remains to be experimentally validated.

In this paper, we compare the End-to-End (E2E) SNER methods using MMSpeech (methods 9 and 10 in Table 2) with the pipeline methods employing MMSpeech for ASR and Bert-large for NER (methods 11 and 12 in Table 2), demonstrating that the E2E approach outperforms the pipeline when using the pretrained MMSpeech model. This suggests that with the pretrained speech model MMSpeech, the E2E approach can mitigate error propagation compared to Pipeline methods and achieve superior performance in SNER. The comparison of End-to-End and Pipeline methods' performance on the SNER task between existing work and this study is presented in Table 6.

G Evaluating the Error Reduction in E2E SNER Methods

To evaluate how the End-to-End (E2E) SNER method reduces error propagation compared to pipeline methods, we designed experiments to test the ASR performance of the E2E SNER approach discussed in this paper, with the experimental results presented in Table 7. The results indicate that the SNER method, trained end-to-end using MMSpeech, shows improved ASR performance over the original MMSpeech model. We believe this improvement is one of the reasons why the E2E

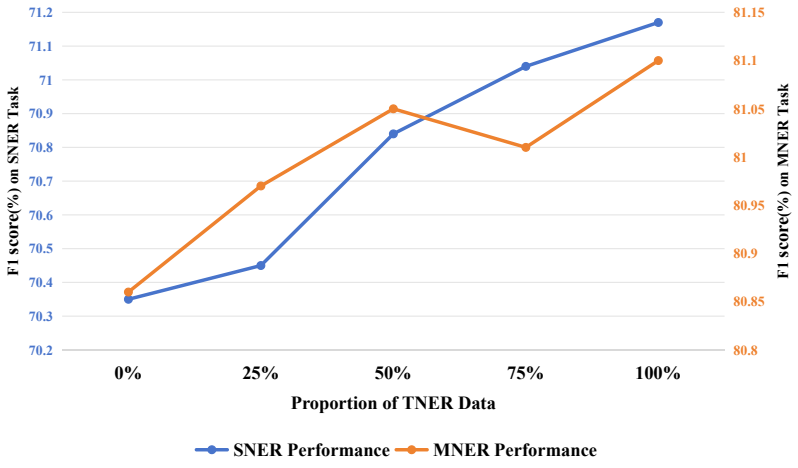
Method	CER(%)	
	ASR	EA-ASR
MMSpeech-base(original)	2.62	-
MMSpeech-large(original)	2.14	-
MMSpeech-base-ETG(SNER)	2.64	2.71
MMSpeech-large-ETG(SNER)	2.10	2.17
IMAGE(MMSpeech-base, only SNER input)	2.59	2.67
IMAGE(MMSpeech-large, only SNER input)	2.11	2.15

Table 7: Performance of Models in Speech Recognition (ASR) for the SNER Task. Models were trained using the AISHELL-NER dataset and evaluated based on Character Error Rate (CER) for ASR and Entity-Aware ASR (EA-ASR). In ASR evaluations, special entity annotation characters are excluded, whereas in EA-ASR, CER calculations include these special characters.

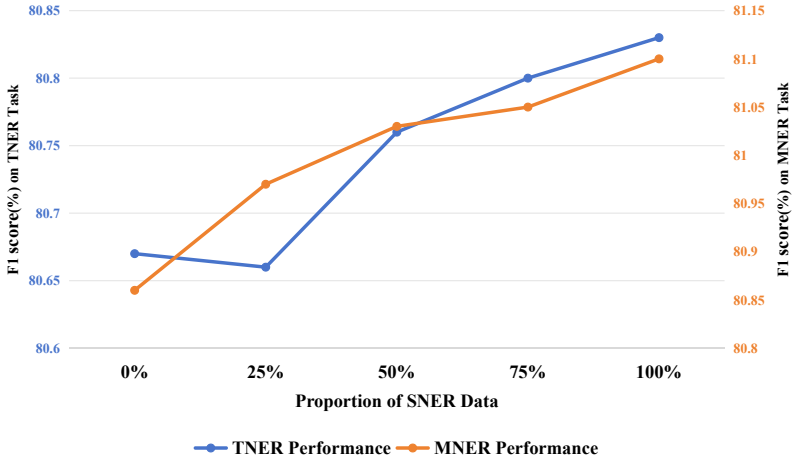
method outperforms the Pipeline method in our study. Additionally, when special characters used to denote entities such as [], (), and <> are included, the calculated Character Error Rate (CER) tends to be higher. This suggests that accurately generating entity annotation symbols in an Entity-Aware text generation task is more challenging than generating the text itself.

H Case Studies

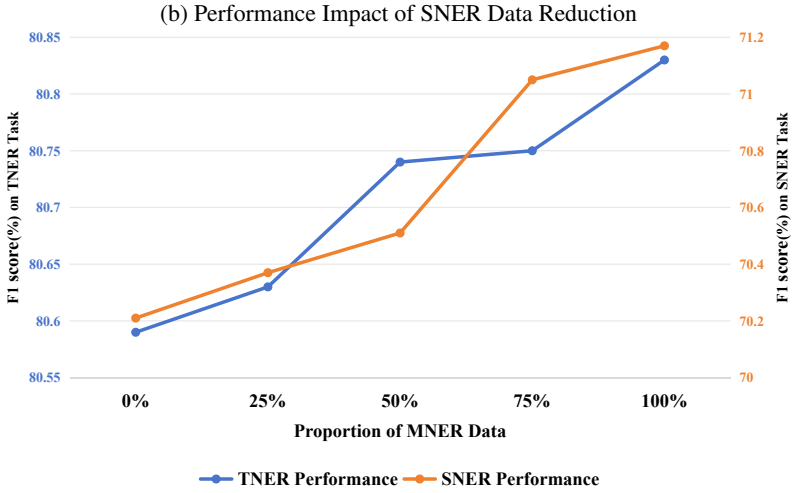
To further reveal how inputs from one modality assist those from another, we have selected two typical cases of the IMAGE framework under different modal inputs, which are displayed in Figure 5. In Case #1, the person entity "孙杨(Sun Yang)" is incorrectly annotated as two separate person entities "孙(Sun)" and "杨(Yang)" when speech modality input is not provided. However, when speech input is available, the pauses within the speech provide strong cues, leading to "孙杨(Sun Yang)" being more likely annotated as a single entity rather than incorrectly as two. This example illustrates how the speech modality can positively assist the text modality. In Case #2, if text input is not provided and only speech input is relied upon, the person entity "李晓霞(Li Xiaoxia)" is mistakenly generated as "李小霞(Li Xiaoxia)", which consists of homophones. This situation frequently occurs in the generation of person and location entities in Chinese SNER. However, if text input is provided at this time, the person entity "李晓霞(Li Xiaoxia)" is correctly identified. This example demonstrates how the text modality can positively assist the speech modality.



(a) Performance Impact of TNER Data Reduction



(b) Performance Impact of SNER Data Reduction



(c) Performance Impact of MNER Data Reduction

Figure 4: Impact of Training Data Ratio Changes on the Performance of IMAGE (MMSpeech-large) on the CNERTA Dataset.

Figure 5: Case studies illustrating how text modality inputs and speech modality inputs mutually assist each other.