Contrastive Token-Wise Meta-Learning for Unseen Performer Visual Temporal-Aligned Translation

Linjun Li*, **Tao Jin***, **Xize Cheng***,
Ye Wang, **Wang Lin**, **Rongjie Huang**, **Zhou Zhao**†
Zhejiang University
{lilinjun21,jint_zju,chengxize}@zju.edu.cn
{yew,linwanglw,rongjiehuang,zhaozhou}@zju.edu.cn

Abstract

Visual temporal-aligned translation aims to transform the visual sequence into natural words, including important applicable tasks such as lipreading and fingerspelling recognition. However, various performance habits of specific words by different speakers or signers can lead to visual ambiguity, which has become a major obstacle to the development of current methods. Considering the constraints above, the generalization ability of the translation system is supposed to be further explored through the evaluation results on unseen performers. In this paper, we develop a novel generalizable framework named **C**ontrastive **To**ken-Wise **M**eta-**l**earning (CtoML), which strives to transfer recognition skills to unseen performers. To the best of our knowledge, employing meta-learning methods directly in the image domain poses two main challenges, and we propose corresponding strategies. First, sequence prediction in visual temporal-aligned translation, which aims to generate multiple words autoregressively, is different from the vanilla classification. Thus, we devise the token-wise diversity-aware weights for the meta-train stage, which encourages the model to make efforts on those ambiguously recognized tokens. Second, considering the consistency of word-visual prototypes across different domains, we develop two complementary global and local contrastive losses to maintain inter-class relationships and promote domain-independence. We conduct extensive experiments on the widely-used lipreading dataset GRID and the fingerspelling dataset ChicagoFSWild, and the experimental results show the effectiveness of our proposed CtoML over existing state-of-the-art methods.

1 Introduction

Human communication is dominated by speech, which conveys semantic information through the acoustic signals. However, persons with dysphonia necessitate reliance on visual perception for independent expressions, such as lip movements and hand gestures. Hence, automatic visual language translation is helpful in bridging the gap between people who communicate through diverse senses.

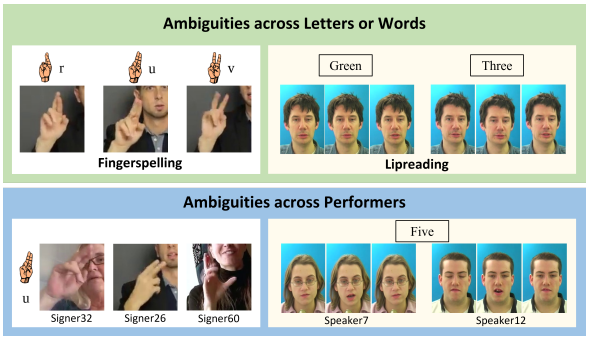


Figure 1: Ambiguities in visual temporal-aligned translation. Green box: Similar performance of the same performer on different words. Blue box: Diverse performance of different performers on the same text unit.

In visual language translation tasks, whether it is observing lip movement or understanding gesture sequences, the common denominator is that the visual content and the translated natural language words are temporally aligned. In this paper, we collectively refer to tasks with the above properties as visual temporal-aligned translation. Specifically, lipreading, *a.k.a.* visual speech recognition (VSR), aims to recognize spoken sentences based on lip movements. Another task is fingerspelling recognition, where recognized text is generated letter-by-letter from the fast and coherent indistinguishable handshapes of signers (Figure 2). Accordingly, mainstream research methods utilize autoregressive models to generate multiple words.

However, current methods (Afouras et al., 2022; Shi et al., 2019) show weakness when applied to real-life scenarios, due to the fact that different performers have a variety of performance habits on specific words leading to ambiguity, as shown in

* Equal contribution.
† Corresponding author

Findings of the Association for Computational Linguistics: ACL 2023, pages 10993–11007

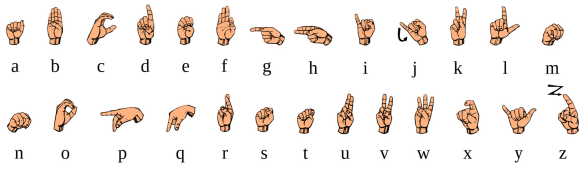


Figure 2: The ASL fingerspelling alphabet, modified from (English Wiki., 2022).

Figure 1. Moreover, the gap between performers can be magnified in data-limited settings, *i.e.* some environments or low-resource languages where collection and annotation are expensive. Ideally, an applicable visual temporal-aligned translation system is supposed to have excellent generalization ability and convincing recognition accuracy for unseen performers. We argue that referring to methods in the domain generalization task to deal with this dilemma is a feasible solution. Concretely, performers with different performance styles or presentation habits can be treated as different domains. For example, in lipreading, some speakers have a slight lip movement, while others have a relatively exaggerated display. Another instance in fingerspelling recognition is that the handshapes of each signer on the same letter are diverse, which could be due to factors such as personal habits and movement directions. Therefore, domain-independent visual temporal-aligned translation can break through the above obstacles, especially on datasets with limited labeled samples.

In this paper, we propose an innovative generalizable framework to deal with the challenging domain-independent visual temporal-aligned translation, called Contrastive Token-Wise Meta-learning (CtoML). Video-sentence pairs of specific performers are used for training, and then we test on unseen performers. As far as we know, directly transferring meta-learning methods in the image domain to lipreading and fingerspelling recognition raises the following two challenges:

First, sequence prediction in visual temporal-aligned translation autoregressively generates multiple words, which is different from vanilla classification. Consequently, we design the token-wise diversity-aware weights for the meta-train phase. The variance of the interacted attention map between performers is measured, and then regarded as the learning difficulty coefficient of the token, so as to concentrate on tackling ambiguous words.

Second, taking into account the consistency of word-visual prototypes across different domains, we develop two complementary contrastive losses. Globally, we frame-word-align the decoded interaction matrix between movement features and sequential sentences to maintain a consistent semantic space of the same class across domains, regardless of the domain-specific variations and class-specific vocabulary positions in the sequence. Locally, relying on contrastive constraints, we facilitate the model to draw closer decoded outputs of words that are semantically similar regardless of domain, and pull away those words that are disparate.

In summary, the token-wise diversity-aware weights we devised encourage the model to focus on tokens with inherent ambiguity in sequence prediction, and the meta-learning process can simulate various domain shift scenarios to assist in finding generalization learning directions. The effectiveness of our CtoML is demonstrated through extensive experiments on the lipreading benchmark dataset GRID and fingerspelling dataset ChicagoFSWild. Our main contributions are as follows:

- We are dedicated to enhancing the generalization ability of the translation system to out-of-domain performers, and correspondingly propose contrastive token-wise meta-learning (CtoML) framework to clarify the generalization learning direction in sequence prediction.
- To focus on the inherently ambiguous words that are confusing to recognize, we devised the token-wise diversity-aware weights to reflect the learning difficulty coefficient of tokens.
- Based on the contrastive constraints, two complementary global and local losses are developed to preserve the inter-class semantic relationships and promote domain-independence.

2 Related Work

2.1 Lip Reading

Lipreading is the task of recognizing spoken sentences from a silent talking face video. Early works are carried out on word-level recognition (Chung and Zisserman, 2016; Wand et al., 2016), and then with the adoption of models developed from ASR tasks, the researchers turn to sentence-level prediction (Assael et al., 2016; Chung et al., 2017; Zhang et al., 2019). Existing studies are primarily based on CTC methods (Assael et al., 2016;

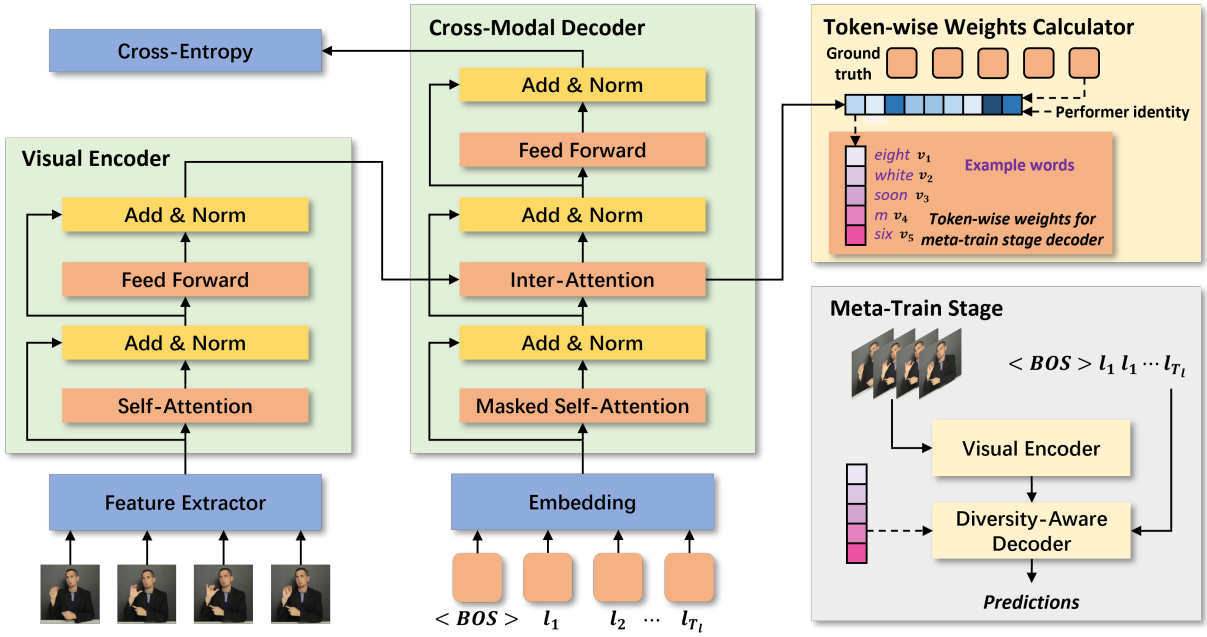


Figure 3: The overall framework of diversity-aware Transformer, composed of visual encoder, token-wise weights calculator and decoder. The diversity-aware decoding with token-wise weights is applied in the meta-train stage.

Petridis et al., 2018; Chen et al., 2020) and autoregressive sequence-to-sequence models (Chung et al., 2017; Zhang et al., 2019; Afouras et al., 2022). Strikingly, Transformer-based architectures (Afouras et al., 2022; Lin et al., 2021) are commonly developed and lead to significant improvements. Distilling knowledge from speech recognition to enhance visual modality in lipreading (Zhao et al., 2020; Ma et al., 2021a) also deserves attention. Additionally, advances from self-supervised representation learning methods (Ma et al., 2021b; Shi et al., 2022; Pan et al., 2022; Cheng et al., 2023) employing pre-training strategies are instructive for visual speech recognition. However, the methods mentioned above do not delve into the generalization ability of lipreading models, which motivates us to explore the direction of model learning on unseen speakers.

2.2 Fingerspelling Recognition

Fingerspelling recognition is a component of sign language recognition that aims to discriminate the fine-grained handshapes of signers. Since the introduction of prior end-to-end models (Koller et al., 2017; Shi and Livescu, 2017; Papadimitriou and Potamianos, 2019) to continuous sign language recognition (Jin et al., 2022b,a), fingerspelling recognition in the wild has achieved substantial progress with a greater emphasis on real-life scenarios (Joze and Koller, 2018; Shi et al., 2018, 2019; Gajurel et al., 2021; S et al., 2021). Subsequently, a multi-task learning manner is proposed in recent researches (Shi et al., 2021; Jiang et al., 2022) to detect effective gesture regions while recognizing fingerspelling signs. Although fingerspelling recognition is a constrained task, it is actually more suitable for evaluating the generalization capability of the model on unseen signers, due to the indistinguishable ambiguity caused by faster fine-grained finger movements.

2.3 Domain Generalization

Domain generalization aims to train a model with limited source domains to generalize directly to unseen target domains. Recently proposed methods are progressive in three aspects, including (1) Representation learning: domain-alignment based method (Mahajan et al., 2021) for learning domain-agnostic representations; disentangled representation method (Zhang et al., 2022) for separating domain-specific and domain-invariant features. (2) Data manipulation: augmentation methods (Shankar et al., 2018; Zhou et al., 2021) to enhance model robustness. (3) Learning strategies: meta-learning (Shu et al., 2021; Jin and Zhao, 2021) with meta-train and meta-test two stages. Despite advancements in speech fusion and synthesis (Huang et al., 2022, 2023; Li et al., 2023), domain generalization on visual temporal-aligned translation tasks has only made preliminary obser-

vations on lipreading from unseen speakers (Assael et al., 2016). For fingerspelling, although the exsiting real-life dataset (Shi et al., 2019) is non-overlapping, the recognition accuracy still retains a large capacity for improvement. Thus, we propose a novel framework called CtoML to solve this task.

3 Approach

3.1 Problem Formulation

We first introduce the problem formulation of visual temporal-aligned translation. Given a sequence of frames from video segments $\boldsymbol{S} = [s_1, s_2, \cdots, s_{T_s}]$, our goal is to predict the textual sequence $\boldsymbol{L} = [l_1, l_2, \cdots, l_{T_l}]$, where s_i is the i-th frame, T_s is the number of frames in the segment, l_j is the j-th word or letter, T_l is the number of transcribed units and $T_l \leq T_s$. Here, the visual content of the video segment $\boldsymbol{S}$ is temporally aligned with the semantics in the textual sequence $\boldsymbol{L}$. In the domain-independent setting, we treat each speaker or clustered signers as a domain, denoted as $D_k = \{(\boldsymbol{S}_n^{(k)}, \boldsymbol{L}_n^{(k)})\}_{n=1}^{N_k}$, where N_k is the number of video-sentence pairs in the k-th domain. The entire K domains are $\mathcal{D} = [D_1, D_2, \cdots, D_K]$. Next, The source training set $\mathcal{D}_{sr}$ and the target testing set $\mathcal{D}_{tg}$ are divided strictly according to the performers, ensuring that the performers appearing in $\mathcal{D}_{sr}$ are not permitted to be seen in $\mathcal{D}_{tg}$, *i.e.* $\mathcal{D}_{sr} \cap \mathcal{D}_{tg} = \varnothing$. Thus, the training set containing video-sentence pairs can be denoted as $\mathbf{T_{sr}} = \{\boldsymbol{S_m}, \boldsymbol{L_m} \mid m \in [1, N_{sr}]\}$, and the testing set as $\mathbf{T_{tg}} = \{\boldsymbol{S_m}, \boldsymbol{L_m} \mid m \in [N_{sr}+1, N_{sr}+N_{tg}]\}$, where N_{sr} and N_{tg} are the numbers of training and testing sets respectively.

3.2 Diversity-Aware Transformer

To enforce the model to concentrate on ambiguously recognized words, we introduce a diversity-aware Transformer with token-wise weights, as shown in Figure 3. Concretely, we devise a token-wise module to capture ambiguities between visual representations of different performers' output from the visual encoder. Integrating the token-wise difficulty coefficient, natural language words are sequentially generated in the meta-train stage.

Visual Encoder. Following vanilla Transformer (Vaswani et al., 2017) and autoregressive visual speech recognition model TM-seq2seq (Afouras et al., 2022), the encoder of diversity-aware Transformer is composed of stacked multi-head self-attention blocks and feed-forward layers. In advance, we prepare the features extracted by the pre-trained model, denoted as $\boldsymbol{F} \in \mathbb{R}^{T_s \times d}$. Then, we can obtain the encoded representations F' through the visual encoder (VisEncoder) as follows:

$$F' = \text{VisEncoder}(F, F, F), \tag{1}$$

where $F' \in \mathbb{R}^{T_s \times d}$. Illustratively, the details of the encoder are provided in Appendix A.

Cross-Modal Decoder. We train a stable task model with the vanilla decoder before the meta-learning phase for subsequent token-wise weight calculations. The standard cross-modal decoder interacts target word embeddings $E \in \mathbb{R}^{T_l \times d}$ with encoded visual features F' to generate character probabilities. Specifically, the decoder is stacked with self-attention, inter-attention, and feed-forward layers. At each time step t, the word embedding ($E_t \in \mathbb{R}^{T_t \times d}$) before t is updated to E'_t via the self-attention layer, as below:

$$E'_t = \text{LN}(E_t + \text{SA}(E_t)), \tag{2}$$

where $E'_t \in \mathbb{R}^{T_t \times d}$, T_t is the word embedding length up to time t, $\text{LN}(\cdot)$ and $\text{SA}(\cdot)$ denote the layer normalization and the self-attention layer, respectively. E'_t is then used as a query to calculate the output I_t of the inter-attention layer. The process with the residual connection is as follows:

$$\begin{aligned} I_t &= \text{LN}(E'_t + \text{MHA}(E'_t, F', F')), \\ I'_t &= \text{LN}(I_t + \text{FFN}(I_t)), \end{aligned} \tag{3}$$

where $I_t, I'_t \in \mathbb{R}^{T_t \times d}$ and I'_t denotes the output of feed-forward network $\text{FFN}(\cdot)$, $\text{MHA}(\cdot)$ denotes multi-head attention. Then, we can produce the probability distribution p_t and give the cross-entropy loss function $\mathcal{L}_{ta}$ as follows:

$$\begin{aligned} p_t &= \text{softmax}(W_p I'_t + b_p), \\ \mathcal{L}_{ta} &= -\sum_{t=1}^{T_l} log p_t(l_t | l_{<t}, F), \end{aligned} \tag{4}$$

where W_p and b_p are trainable parameter matrices.

Token-wise Weights Calculator. To obtain the learning difficulty coefficient of tokens, we compute the diversity-aware weights using the attention maps of the inter-attention layers between different performers. We denote the final layer output of $\text{MHA}(E'_t, F', F')$ in Eqn.(3) as $U_t = [u_r]_{r=1}^{T_t}$, and

$u_r \in \mathbb{R}^d, U_t \in \mathbb{R}^{T_t \times d}$. Next, the identity performance $a_c^{(k)} \in \mathbb{R}^d$ of the k-th performer on word c can be denoted as:

$$a_c^{(k)} = \frac{1}{N_k^{(c)}} \sum_{r:l_r^{(k)}=c} (u_r^{(k)}), \tag{5}$$

where $u_r^{(k)}$ is the attention vector of the k-th performer, and $N_k^{(c)}$ is the number of the samples labelled as c. Subsequently, we compute the variance of each word across different performers to reflect ambiguities due to various motion habits, given by:

$$v_c = \sigma(\frac{1}{K}\sum_{k=1}^{K}(a_c^{(k)})^2 - (\frac{1}{K}\sum_{k=1}^{K}a_c^{(k)})^2), \tag{6}$$

where $v_c \in \mathbb{R}^d$, c is the index of the word in the vocabulary, σ denotes the non-linear activation function such as Sigmoid. Hence, complete word difficulty representations $V = [v_c]_{c=1}^{T_c} \in \mathbb{R}^{T_c \times d}$ are produced, which are provided to the diversity-aware decoding. T_c is the vocabulary length.

Diversity-aware Decoding. After requiring the token-wise weights, we perform diversity-aware decoding on the encoded features in the meta-train stage. Specifically, we apply the token-wise weights to the cross-entropy loss in Eqn.(4) and obtain a developed loss $\mathcal{L}_{da}$ given by:

$$\mathcal{L}_{da} = -\sum_{t=1}^{T_l} V * logp_t(l_t|l_{<t}, F), \tag{7}$$

where $*$ denotes the token-wise multiplication of vectors. Therefore, our model can consciously focus on ambiguous words under the adjustment of the token-wise diversity-aware weights.

3.3 Contrastive Meta-learning

Our methods train the task model on meta-train sets and then improve the generalization ability on meta-test sets. Furthermore, inter-class semantic relationships are consolidated while promoting domain-independence under two complementary contrast constraints. During the meta-train stage, our prior computed token-wise weights can cooperate with the specific task loss. The complete meta-learning process is summarized in Alg.1.

Concretely, we randomly split entire K domains $\mathcal{D}_{sr}$ into meta-train ($\mathcal{D}_{tr} = \{D_i\}_{i=1}^{N_{tr}}$) and meta-test ($\mathcal{D}_{te} = \{D_j\}_{j=1}^{N_{te}}$) domains in each epoch, where $N_{te} = N_{sr} - N_{tr}$. During the meta-train stage, the parameters are updated from the task-supervised cross-entropy loss function $\mathcal{L}_{da}$, calculated as $\theta' = \theta - \alpha\nabla_\theta\mathcal{L}_{da}(\mathcal{D}_{tr};\theta)$. θ denotes all the trainable parameters and α is the learning rate. Then, the relationship between ambiguous words and the semantic space consistencies among performers are preserved and harmonized in the following meta-test stage.

Algorithm 1 Contrastive Meta Visual Temporal-Aligned Translation

Input: Source training domains $\mathcal{D}_{sr} = \{D_m\}_m^{N_{sr}}$,
Initialize: Model parameters θ; hyperparameters α, β, λ

while not converged **do**
  Randomly split $\mathcal{D}_{sr}$ into $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$
  Sample a batch B_{tr} from $\mathcal{D}_{tr}$ ▷ Meta-train
  for all B_{tr} **do**
    Compute task-specific loss $\mathcal{L}_{da}(B_{tr};\theta)$
    $\theta' \leftarrow \theta - \alpha\nabla_\theta\mathcal{L}_{da}$
  end for
  Sample a batch B_{te} from $\mathcal{D}_{te}$ ▷ Meta-test
  for all B_{te} **do**
    Compute global loss $(B_i \in B_{tr}, B_j \in B_{te})$: $\mathcal{L}_{gl}(B_i, B_j; \theta')$
    Compute local loss $(B_{sr} \leftarrow [B_{tr}, B_{te}])$: $\mathcal{L}_{lc}(B_{sr};\theta')$
    $\mathcal{L}_{obj} \leftarrow \mathcal{L}_{da} + \lambda(\mathcal{L}_{gl} + \mathcal{L}_{lc})$
  end for
  $\theta \leftarrow \theta - \beta\nabla_\theta\mathcal{L}_{obj}$
end while
return Model parameters θ

In detail, we devise two complementary contrast constraints. With a global objective of stabilizing inter-class relationships, we attempt to preserve the relationship between learned ambiguous words on unseen performers. Hence, for a specific performer, we can compute a word distribution $g_c^{(k)}$ with the personality vector obtained by Eqn.(5) and a softmax at temperature τ. Then, the global loss $\mathcal{L}_{gl}$ can be gained by minimizing the symmetrized relative entropy as follows:

$$g_c^{(k)} = \mathrm{softmax}(a_c^{(k)}/\tau), \tag{8}$$

$$\mathcal{L}_{gl} = \frac{1}{2T_c}\sum_{o=1}^{N_o}\sum_{c=1}^{T_c}(H(g_{c,o}^{(i)}\|g_{c,o}^{(j)}) + H(g_{c,o}^{(j)}\|g_{c,o}^{(i)})),$$

where N_o is the number of pairs of (D_i, D_j), $g_{c,o}^{(i)}$ denotes the distribution of D_i in the o-th pair on word c, T_c is the vocabulary length and $H(p\|q) = \sum_z \log(\frac{p_z}{q_z})$ is the details of relative entropy.

Crucially, predictions should not be sensitive to unseen performers, thus requiring a local objective

Method	S(1&2&20&22)		S(3&4&23&24)		S(5&6&25&26)		S(7&8&27&28)	
	WER	CER	WER	CER	WER	CER	WER	CER
Lin et al. 2021	14.71	8.59	12.04	6.65	13.11	7.94	12.88	7.51
Assael et al. 2016	11.40	6.40	9.28	4.32	10.67	5.73	10.41	5.34
Afouras et al. 2022	10.26	5.32	7.81	3.09	9.03	4.26	8.73	4.18
Xu et al. 2018	9.58	5.43	8.20	3.49	9.48	5.01	9.36	5.10
Shi et al. 2022	6.64	3.98	4.98	2.76	6.17	3.59	5.34	3.13
Ours(base)	7.53	4.49	5.47	3.11	6.82	3.98	6.56	3.74
Ours	**5.42**	**2.63**	**3.12**	**1.27**	**4.59**	**2.13**	**4.25**	**2.01**

Table 1: Results of CtoML on the four splits of GRID dataset compared to the baselines and variants. S(1&2&20&22) represents four unseen speakers S1, S2, S20 and S22, and the others are similar. All values here are percentages.

of reducing ambiguous word overlap regardless of performers. We put together $u_r^{(k)}$ in Eqn.(5) of all samples in $\mathcal{D}_{sr}$ in a performer-insensitive manner to obtain a set A of token-wise attention vectors, containing N_z tokens. Next, the contrastive loss $\mathcal{L}_{lc}$ we exploit can be calculated as follows:

$$\mathcal{L}_{lc} = \frac{1}{N_b}\sum_{b=1}^{N_b}(Y * \max(\xi - \rho(x_b, y_b), 0)^2 + (1-Y) * \rho(x_b, y_b)), \quad (9)$$

where $\rho(\cdot)$ denotes a distance function, (x_b, y_b) is the b-th random sample pair in set A, N_b is the number of the sample pairs, $Y = 1 - [x_l = y_l]$, x_l and y_l are the respective labels, margin ξ is to con trol the distance between two samples. Considering the computational complexity, when we sample pairs, we use the sorted queue that dequeues every two elements, instead of enumeration.

We optimize the meta visual temporal-aligned translation model with the developed task loss $\mathcal{L}_{da}$ and contrastive constraints $\mathcal{L}_{gl}$ and $\mathcal{L}_{lc}$, given by:

$$\begin{aligned} \mathcal{L}_{obj} &= \mathcal{L}_{da} + \lambda(\mathcal{L}_{gl} + \mathcal{L}_{lc}), \\ \theta &= \theta - \beta\nabla_\theta \mathcal{L}_{obj}, \end{aligned} \quad (10)$$

where λ is utilized to control the balance, and θ is the learning rate. During the inference stage, we use a common visual encoder and cross-modal decoder without meta stages.

4 Experiments

We evaluate and compare our CtoML on two challenging visual temporal-aligned translation tasks, lipreading and fingerspelling recognition. Our experiments are conducted on two datasets: GRID (Cooke et al., 2006) for lipreading, and ChicagoFSWild (Shi et al., 2018) for fingerspelling recognition. In this section, we provide a brief introduction to the datasets and corresponding evaluation metrics. Then, we present concrete experimental settings and compare CtoML with baseline methods. Subsequently, we analyze the main results and conduct ablation studies. Besides, we also provide qualitative examples and analysis on GRID and ChicagoFSWild dataset in the Appendix D.

4.1 Dataset

GRID: The GRID dataset contains 33,000 sentences uttered by 33 speakers. The vocabulary of the GRID dataset includes 51 different words in 6 categories. For evaluation on unseen speakers, four speakers (s1, s2, s20, s22) are selected by (Assael et al., 2016) as the test set. Similar to the above split, we provide three additional splits to discuss the robustness of the model. The unseen speakers in these three splits are: (s3, s4, s23, s24), (s5, s6, s25, s26) and (s7, s8, s27, s28).

ChicagoFSWild: The ChicagoFSWild dataset contains 7,304 fingerspelling clips performed by 160 signers. The data is split into three sets with no overlapping signers: 5,455 training sentences from 87 signers, 981 development sentences from 37 signers, and 868 test sentences from 36 signers. The vocabulary size is 31 including 26 alphabets and 5 special characters. In this paper, we follow the split in (Shi et al., 2018).

4.2 Experiments for Lipreading

Evaluation Metrics: Following prior works (Assael et al., 2016), we evaluate the performance based on the metrics of character error rate (CER) and word error rate (WER). The error rates can be computed as: $ErrorRate = \frac{(S+D+I)}{M}$, where S, D, I are the number of the substitutions, deletions and insertions in the alignments, and M is the number of characters or words.

Method	S(1&2&20&22)		S(3&4&23&24)		S(5&6&25&26)		S(7&8&27&28)	
	WER(%)	CER(%)	WER(%)	CER(%)	WER(%)	CER(%)	WER(%)	CER(%)
w/o.TAW	5.71	2.86	3.51	1.60	4.88	2.61	4.57	2.39
w/o.CS	5.83	2.92	3.59	1.64	4.95	2.67	4.62	2.45
w/o.Meta	6.06	3.27	3.84	1.93	5.20	2.85	4.94	2.73
Ours	**5.42**	**2.63**	**3.12**	**1.27**	**4.59**	**2.13**	**4.25**	**2.01**

Table 2: Ablation results of our CtoML on GRID dataset.

Method		S(1&2&20&22)		S(7&8&27&28)	
$\mathcal{L}_{gl}$	$\mathcal{L}_{lc}$	WER	CER	WER	CER
-	-	5.83	2.92	4.62	2.45
✓	-	5.66	2.73	4.41	2.28
-	✓	5.72	2.86	4.50	2.37
✓	✓	**5.42**	**2.63**	**4.25**	**2.01**

Table 3: Ablation study of two devised complementary contrastive constraints on GRID dataset.

Method	Letter Accuracy(%)	
	dev	test
HDC-FSR	42.8	41.9
IAF-FSR	46.8	45.1
FGVA	47.0	48.4
TDC-SL	-	50.0
Ours(w.ResNet18)	54.9	54.1
Ours(w.ResNet50)	**55.7**	**54.9**

Table 4: Results of CtoML on ChicagoFSWild dataset compared to the baselines, where dev, test denotes the development and test sets, respectively.

Implementation Details: The videos are first processed with the Dlib detector (King, 2009), and then we extract a mouth-centered crop of size 100 $\times$ 60 as the video input. We augment the dataset with horizontal flips with 50% probability. According to (Shi et al., 2022), we obtain robust feature representations for our model. For each meta-train stage, we perform 10 iterations and take the last updated parameter as θ'. Also, more training and parameter settings are listed in the Appendix C.

Results and Analysis: We compare our CtoML with several state-of-the-art methods, LipNet (Assael et al., 2016), SimulLR (Lin et al., 2021), LCANet (Xu et al., 2018), TM-seq2seq (Afouras et al., 2022) and AV-HuBERT(Shi et al., 2022). We denote CtoML without modules designed for the generalization objective as **Ours(base)**, which is trained with only the task loss $\mathcal{L}_{ta}$. Table 1 summarizes the results of unseen speaker lipreading on GRID dataset with a comparison to baselines. Across all domain splits, our CtoML outperforms the state-of-the-art method AV-HuBERT with an average of 1.44% on WER and 1.36% on CER. In comparison, our method provides token-wise diversity-aware weights, which supports the model to grasp the learning direction of ambiguous words better. Furthermore, we exploit the essence of domain generalization to improve the model's generalization capability to unseen speakers, which has not been considered in previous methods. Moreover, the performance of **Ours(base)** is comparable to that of the state-of-the-art methods, further demonstrating the effectiveness of our modules for generalization to unseen performers. Notably, SimulLR (Lin et al., 2021), which performs well on overlapping regular split, struggles with domain-independent settings. It suggests that methods with special objectives may not be generalized effectively. In addition, we can see that the evaluation results fluctuate across different divisions, indicating that there are indeed significant differences in performance habits between speakers.

Ablation Study: We conduct extensive ablation studies on token-wise diversity-aware weights, contrastive constraints, and meta-learning strategies to represent all contributions. Table 2 show the capabilities of each key module, where **w/o. TAW** denotes the model without token-wise weights in the diversity-aware decoding, **w/o. CS** denotes the model without contrastive constraints, and **w/o. Meta** denotes the model without meta-learning. We can observe that CtoML performs significantly better than **w/o. Meta**, which demonstrates the effectiveness of meta-learning in improving the generalization capability. As expected, the evaluation results of **w/o. TAW** is mediocre, because the model can get lost in the inherent ambiguous words without the coordination of token-wise weights. Comparing CtoML with **w/o. CS**, we find that CtoML achieves relatively superior results, indicating that contrastive constraints can work smoothly.

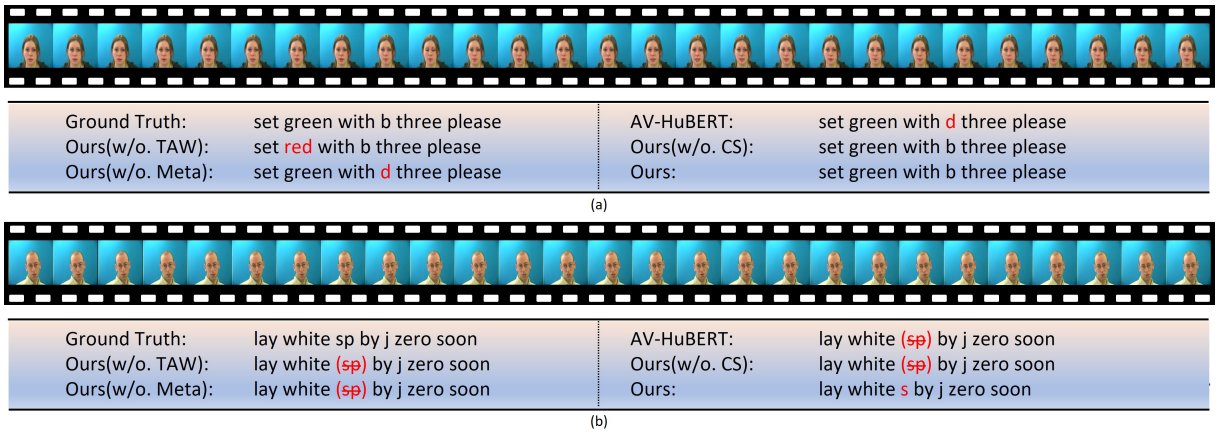


Figure 4: Qualitative results of GRID. Words in red are incorrect predictions.

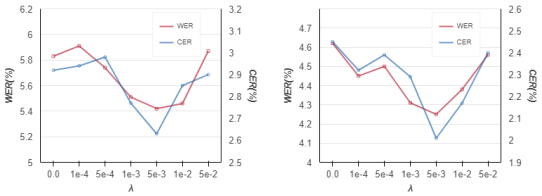
Furthermore, we control the constraints to explore the contribution of a specific loss, as shown in Table 3. Since the two losses are complementary, we can see that the absence of either loss can bring about a considerable decrease. The impacts of λ are shown in Figure 5. It is clear that both WER and CER fluctuate and become poor when λ is too small or too large. The model achieves best when λ is 5×10^{-3}. More observations on hyperparameters are provided in the Appendix B.

Qualitative Analysis: Figure 4 shows the qualitative results on GRID. We provide a comparison of the predicted sentences of CtoML and AV-HuBERT. Intuitively, CtoML performs better due to the joint effect of the devised token-wise diversity-aware wights, complementary contrastive losses and meta-learning strategy. In the first example, we can see that when encountering the letters b and d with similar lip movements, the AV-HuBERT appears weak. In contrast, CtoML effectively guides the direction to deal with ambiguous words and thus predicts the text sequence correctly. Although our model does not successfully predict the second case, it captures fast and ambiguous features and finds tokens similar to the ground truth.

4.3 Experiments for Fingerspelling

Due to space constraints, implementation details and qualitative analysis are put in Appendix C,D.

Evaluation Metrics: For evaluation, the letter accuracy modified from Levenshtein string edit distance $1-\frac{(S+D+I)}{M}$ is adopted (Shi et al., 2019). Here, the sum of S, D, I is the minimum number that transforms the prediction to ground truth, and M is the number of ground-truth letters.



(a) S(1,2,20,22) (b) S(7,8,27,28)

Figure 5: The impact of λ on WER and CER.

Results and Analysis: For the fingerspelling recognition, our CtoML is compared with four state-of-the-art methods, HDC-FSR(Shi et al., 2018), IAF-FSR(Shi et al., 2019), FGVA(Gajurel et al., 2021), and TDC-SL(Papadimitriou and Potamianos, 2020). Illustratively, since the authors do not name their methods, the names used here are abbreviations assigned based on specific characteristics. We use **Ours(w. ResNet18)** to denote the model that adopts ResNet18 rather than ResNet50 (He et al., 2016). From Table 4, we can find that CtoML outperforms the others by averaging at least 4.9% and 8.7% on the development and test set. This is because the performance of signers with the same letter can be diverse, and our complementary contrastive constraints allow models to understand ambiguous words while maintaining a consistent semantic space on unseen signers. Convincingly, **Ours(w. ResNet18)** is significantly higher than the others by a margin of 4.1% on the test set, even though we use a weaker extractor.

5 Conclusion

We have proposed a new framework called contrastive token-wise meta-learning (CtoML) for visual temporal-aligned translation, which promotes

the generalization capability on unseen performers. To concentrate on the inherently ambiguous words, we devise token-wise diversity-aware weights. Furthermore, we develop contrastive meta-learning to clarify the learning direction. Reasonable complementary contrast constraints are provided to preserve inter-class semantic relationships and promote domain-independence. The experimental results on GRID and ChicagoFSWild dataset demonstrate the effectiveness of CtoML.

6 Limitations

In this section, we develop a clear discussion of the limitations of this paper. Our method faces obstacles when attempting to validate it on datasets other than those previously used in this paper. For example, LRS2 (Afouras et al., 2022) is a widely used dataset in visual language recognition tasks. However, since LRS2 dataset does not provide speaker identification labels, we cannot easily classify speakers into domain-specific and domain-independent sets. Despite the enormous amount of work, re-annotating existing datasets with crowdsourcing or annotating a new real-life dataset with speaker labels is a viable solution. Besides, existing data augmentation methods cannot match the generalization requirements on visual temporal-aligned translation perfectly, which inspires researchers to develop targeted augmentation paradigms based on the study in this paper to cooperate with our meta-learning strategies.

7 Ethics Statement

The datasets used in this study were those produced by previous researchers, and we followed all relevant legal and ethical guidelines for their acquisition and use. Furthermore, we recognize the potential moral hazard of visual temporal-aligned translation tasks, such as their use in surveillance or listening. We are committed to conducting our research ethically and ensuring that our research is beneficial.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No. 62222211, Grant No.61836002 and Grant No.62072397.

References

Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2022. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):8717–8727.

Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*.

Weicong Chen, Xu Tan, Yingce Xia, Tao Qin, Yu Wang, and Tie-Yan Liu. 2020. Duallip: A system for joint lip reading and generation. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1985–1993. ACM.

Xize Cheng, Linjun Li, Tao Jin, Rongjie Huang, Wang Lin, Zehan Wang, Huangdai Liu, Ye Wang, Aoxiong Yin, and Zhou Zhao. 2023. Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition. *arXiv preprint arXiv:2303.05309*.

Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3444–3453. IEEE Computer Society.

Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, volume 10112 of *Lecture Notes in Computer Science*, pages 87–103.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

English Wiki. 2022. American sign language. https://en.wikipedia.org/wiki/American_Sign_Language.

Kamala Gajurel, Cuncong Zhong, and Guanghui Wang. 2021. A fine-grained visual attention approach for fingerspelling recognition in the wild. In *2021 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2021, Melbourne, Australia, October 17-20, 2021*, pages 3266–3271. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*.

Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. 2022. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*.

Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. In *Advances in Neural Information Processing Systems*.

Ziqi Jiang, Shengyu Zhang, Siyuan Yao, Wenqiao Zhang, Sihan Zhang, Juncheng Li, Zhou Zhao, and Fei Wu. 2022. Weakly-supervised disentanglement network for video fingerspelling detection. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 5446–5455. ACM.

Tao Jin and Zhou Zhao. 2021. Contrastive disentangled meta-learning for signer-independent sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5065–5073.

Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. 2022a. Mc-slt: Towards low-resource signer-adaptive sign language translation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4939–4947.

Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. 2022b. Prior knowledge and memory enriched transformer for sign language translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3766–3775.

Hamid Reza Vaezi Joze and Oscar Koller. 2018. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*.

Davis E. King. 2009. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Oscar Koller, Sepehr Zargaran, and Hermann Ney. 2017. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3416–3424. IEEE Computer Society.

Linjun Li, Tao Jin, Wang Lin, Hao Jiang, Wenwen Pan, Jian Wang, Shuwen Xiao, Yan Xia, Weihao Jiang, and Zhou Zhao. 2023. Multi-granularity relational attention network for audio-visual question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1.

Zhijie Lin, Zhou Zhao, Haoyuan Li, Jinglin Liu, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2021. Simullr: Simultaneous lip reading transducer with attention-guided adaptive memory. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1359–1367. ACM.

Pingchuan Ma, Brais Martínez, Stavros Petridis, and Maja Pantic. 2021a. Towards practical lipreading with distilled and efficient models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 7608–7612. IEEE.

Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W. Schuller, and Maja Pantic. 2021b. Lira: Learning visual speech representations from audio through self-supervision. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3011–3015. ISCA.

Divyat Mahajan, Shruti Tople, and Amit Sharma. 2021. Domain generalization using causal matching. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7313–7324. PMLR.

Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. 2022. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4491–4503. Association for Computational Linguistics.

Katerina Papadimitriou and Gerasimos Potamianos. 2019. End-to-end convolutional sequence learning for ASL fingerspelling recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2315–2319. ISCA.

Katerina Papadimitriou and Gerasimos Potamianos. 2020. Multimodal sign language recognition via temporal deformable convolutional sequence learning. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2752–2756. ISCA.

Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. 2018. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens,*

Greece, December 18-21, 2018, pages 513–520. IEEE.

Srinivas Kruthiventi S. S, George Jose, Nitya Tandon, Rajesh Roshan Biswal, and Aashish Kumar. 2021. Fingerspelling recognition in the wild with fixed-query based visual attention. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4362–4370. ACM.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. 2018. Generalizing across domains via cross-gradient training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2021. Fingerspelling detection in american sign language. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4166–4175. Computer Vision Foundation / IEEE.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Bowen Shi and Karen Livescu. 2017. Multitask training with unlabeled data for end-to-end sign language fingerspelling recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*, pages 389–396. IEEE.

Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2019. Fingerspelling recognition in the wild with iterative visual attention. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5399–5408. IEEE.

Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Jonathan Michaux, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2018. American sign language fingerspelling recognition in the wild. In *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*, pages 145–152. IEEE.

Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. 2021. Open domain generalization with domain-augmented meta-learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9624–9633. Computer Vision Foundation / IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Michael Wand, Jan Koutník, and Jürgen Schmidhuber. 2016. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 6115–6119. IEEE.

Kai Xu, Dawei Li, Nick Cassimatis, and Xiaolong Wang. 2018. Lcanet: End-to-end lipreading with cascaded attention-ctc. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 548–555. IEEE Computer Society.

Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P. Xing. 2022. Towards principled disentanglement for domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8014–8024. IEEE.

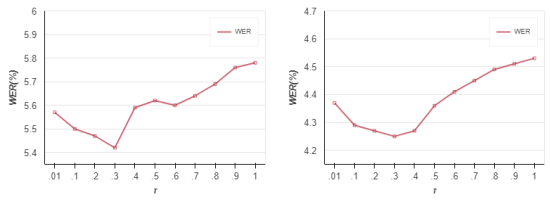
Xingxuan Zhang, Feng Cheng, and Shilin Wang. 2019. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 713–722. IEEE.

Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. 2020. Hearing lips: Improving lip reading by distilling speech recognizers. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6917–6924. AAAI Press.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2021. Domain generalization with mixstyle. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Appendix

In this Appendix, we present the concrete structure of the visual encoder in Section A. Then, the impact of the important hyperparameters is shown in Section B. In addition, we provide more implementation details in Section C. At last, we supplement additional qualitative results and analysis in Section D.



(a) S(1,2,20,22) (b) S(7,8,27,28)

Figure 6: The impact of temperature τ.

A Encoder Details

In this section, we provide the concrete structure of the visual encoder. With prepared visual features, the self-attention layer (SA) is shown as follows:

$$\mathrm{SA}(F) = \mathrm{MultiHead}(F, F, F), \tag{11}$$

where $\mathrm{MultiHead}(\cdot)$ denotes multi-head attention that projects randomly initialized matrices into different representation subspaces. The multi-head attention can be calculated by multiple single heads:

$$\begin{aligned} \mathrm{MHA}(F,F,F) &= \mathrm{Concat}(h_1, h_2, \cdots, h_h)W_1, \\ h_i &= \mathrm{ATT}(FW_i^Q, FW_i^K, FW_i^V). \end{aligned} \tag{12}$$

Here, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times\frac{d}{h}}$ and $W_1 \in \mathbb{R}^{d\times d}$ are trainable parameter matrices, h_i denotes the i-th head and h is the number of heads. $\mathrm{MHA}(\cdot)$ is the short form of multi-head attention and $\mathrm{ATT}(\cdot)$ represents scaled dot-product attention as follows:

$$\mathrm{ATT}(Q, K, V) = \mathrm{softmax}(\frac{Q \cdot K^\top}{\sqrt{d_k}}), \tag{13}$$

where Q (similar to K, V) denotes FW_i^Q (similar to FW_i^K, FW_i^V) and d_k is the dimension of matrix K. A residual connection and layer normalization $\mathrm{LN}(\cdot)$ are followed after each self-attention layer:

$$X = \mathrm{LN}(F + \mathrm{SA}(F)). \tag{14}$$

Subsequently, incorporating a feed-forward network $\mathrm{FFN}(\cdot)$ with transformation layers and a nonlinear activation function σ, we can obtain the encoded features F′ as:

$$\mathrm{FFN}(X) = W_3\sigma(W_2X), \tag{15}$$

$$F' = \mathrm{LN}(X + \mathrm{FFN}(X)), \tag{16}$$

where $W_2 \in \mathbb{R}^{4d\times d}$, $W_3 \in \mathbb{R}^{d\times 4d}$ are trainable weight matrices and $F' \in \mathbb{R}^{T_s\times d}$.

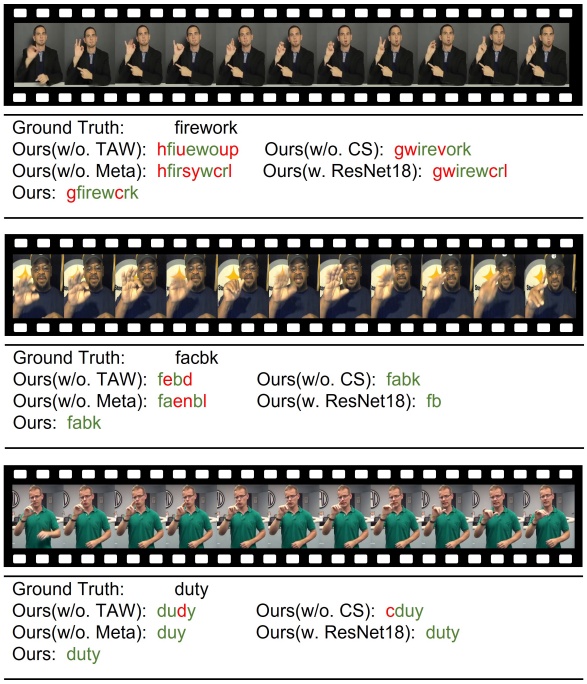


Figure 7: Qualitative results of ChicagoFSWild. The green characters are predictions that meet our expectations, while the red characters are wrong predictions or caused the evaluation metric to drop.

Method	WER(%)	CER(%)
$\alpha=5\times10^{-4}, \beta=5\times10^{-4}$	5.48	2.75
$\alpha=1\times10^{-3}, \beta=1\times10^{-3}$	5.60	2.86
$\alpha=5\times10^{-4}, \beta=1\times10^{-3}$	5.51	2.72
Ours	**5.42**	**2.63**

Table 5: The impact of different initialization learning rates on GRID S(1&2&20&22).

B Hyperparameter Analysis

As depicted in Table 5, we tune the initialization combination of learning rates α and β and determine that $\alpha = 1\times10^{-3}, \beta = 5\times10^{-4}$ has the best performance. Besides, the result in Figure 6 shows that when τ is 0.2, the performance achieves the best on both partitions.

C Implementation Details

Lipreading: In terms of training details, we set the hidden size d to 512 for GRID. The number of heads and attention blocks in multi-head attention mechanisms is 8 and 3, respectively. The dropout rate is set to 0.3. We optimize the loss function using the Adam optimizer (Kingma and Ba, 2014) with learning rates α, β initialized to 1×10^{-3} and 5×10^{-4}. The coefficient λ in $\mathcal{L}_{obj}$ is set to 0.005. The maximum number of epochs is 30

and the batch size is 32 with an NVIDIA GeForce RTX 3090 GPU. For our CtoML, each epoch takes around 3 hours.

Fingerspelling For each video segment, we use a face detector to gain a face-centered crop, which is consistent with (Shi et al., 2019). We use ResNet50 (He et al., 2016) that is pre-trained on ImageNet (Deng et al., 2009) to extract features for sampled resized frames of 112×112. The hidden size d is set to 512, the number of heads is 8. We set the dropout rate to 0.2. As for attention blocks, we use 3 in both the encoder and decoder. Adam algorithm is selected to optimize and the learning rates α, β are initialized to 5×10^{-4}. λ that controls the balance of the loss function is set to 0.005. The batch size is 32 and the maximum number of epochs is set to 30. As for the setting of the meta-train stage, it is the same as lipreading.

D Qualitative Results

In this section, we provide several examples to conduct qualitative analysis on dataset ChicagoFSWild to prove the superiority of our proposed CtoML. From Figure 7, we can find that our CtoML predicts the most promising results with all the proposed modules. The green characters are predictions that meet our expectations, while the red characters are wrong predictions or caused the evaluation metric to drop. The three examples from top to bottom are representative samples selected from the train, development, and test sets. Note that the top example is the result of the training process, only to observe the impact of ambiguous words, and does not reflect the effect of the model. Through the above examples, we can draw the following three facts: (1) The visual performance of some characters is too similar to be the main factor that confuses the model. (2) In the continuous signs, the characters in the middle will be more likely to cause prediction errors or omissions due to too fast or incomplete sign. (3) Due to the production of the dataset, irrelevant signs may appear at the beginning of the video clip, resulting in redundant predicted characters. For the first two challenges, our model shows considerable performance due to the excellent generalization ability.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
See Section 6.

☑ A2. Did you discuss any potential risks of your work?
See Section 7.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
See Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

See Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
See Appendix C.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 4.2 and Appendix B.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
See Appendix C.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
See Section 4.2, 4.3 and Appendix C.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.