TAVT:Towards Transferable Audio-Visual Text Generation

Wang Lin[*], **Tao Jin**[*], **Ye Wang**[*],
Wenwen Pan, **Linjun Li**, **Xize Cheng**, **Zhou Zhao**[†]
Zhejiang University
{linwanglw,jint_zju,yew}@zju.edu.cn
{wenwenpan,lilinjun21,chengxize,zhaozhou}@zju.edu.cn

Abstract

Audio-visual text generation aims to understand multi-modality contents and translate them into texts. Although various transfer learning techniques of text generation have been proposed, they focused on uni-modal analysis (*e.g.,* text-to-text, visual-to-text) and lack consideration of multi-modal content and cross-modal relation. Motivated by the fact that humans can recognize the timbre of the same low-level concepts (*e.g.,* footstep, rainfall, and laughing), even in different visual conditions, we aim to mitigate the domain discrepancies by audio-visual correlation. In this paper, we propose a novel Transferable Audio-Visual Text Generation framework, named TAVT, which consists of two key components: Audio-Visual Meta-Mapper (AVMM) and Dual Counterfactual Contrastive Learning (DCCL). (1) AVMM first introduces a universal auditory semantic space and drifts the domain-invariant low-level concepts into visual prefixes. Then the reconstruct-based learning encourages the AVMM to learn "which pixels belong to the same sound" and achieve audio-enhanced visual prefix. The well-trained AVMM can be further applied to uni-modal setting. (2) Furthermore, DCCL leverages the destructive counterfactual transformations to provide cross-modal constraints for AVMM from the perspective of feature distribution and text generation. (3) The experimental results show that TAVT outperforms the state-of-the-art methods across multiple domains (cross-datasets, cross-categories) and various modal settings (uni-modal, multi-modal).

1 Introduction

Audio-visual text generation bridges the gap between perception (visual and auditory) and communication (via language), and is hence becoming an increasingly important goal for artificial agents. Uni-modal text generation tasks like machine translation (Wang et al., 2022; Jin et al., 2022b; Yin et al., 2022), and image caption (Chen et al., 2017; Tewel et al., 2021; Hu et al., 2022) have already flourished as a result of the large-scale pre-training and huge model capacity. However, for audio-visual text generation tasks, data annotation is more arduous (temporal structure) and expensive (requires monitors and speakers) than uni-modal text generation. Moreover, despite the effectiveness, existing works (Iashin and Rahtu, 2020a; Le et al., 2020; Hori et al., 2021) inevitably suffer severe degradation due to varying construction conditions in different domains.

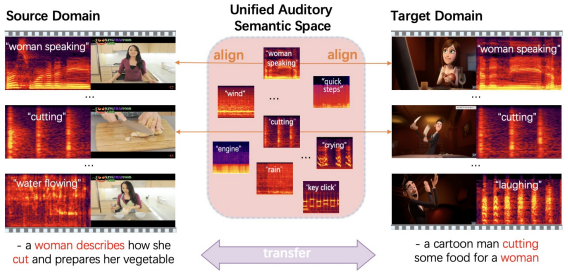

Figure 1: Examples of transferable multi-modal text generation in the source domain (real-word) and the target domain (animation). For the same events ("speaking", "cutting"), although the visual differences are significant, the sounds are similar.

In this paper, to break through the constraint, we propose a novel task, named transferable audio-visual text generation. The main challenge of this task is the multi-modal domain shifts caused by varying conditions, like the visual style and audio energy. One common approach to handle the domain shift is domain-alignment-based transfer learning. However, existing works (Sun and Saenko, 2016; Rozantsev et al., 2018; Ding et al., 2022) focus on the uni-modal analysis, which is insufficient due to the lack of consideration of cross-modal relations.

We observe that while audio and visual are often correlated in natural events and jointly affect human perception, they have different characteristics. As shown in Figure 1, the fact *Timbre is an*

[*] Equal contribution.
[†] Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14983–14999

intrinsic property of the object leads to sounds of the same concept ("speaking" or "cutting") being similar across domains where the appearance like background, perspective, and style is significantly different. Based on this phenomenon, domain invariant low-level concepts can be extracted from the visual with the supervision of the audio, which is pervasive, reliable, and cheaper in contrast to expensive human annotation.

Grounded on the above discussions, we propose an Audio-Visual Meta-Mapper network (AVMM). The key idea of AVMM is to use a universal auditory semantic space to align low-level concepts across different visual domains. In particular, we introduce a visual prefix that serves as a multi-modal bridge between the visual and the audio. Then, we reconstruct audio features from the universal auditory semantic space and produce the audio-enhanced visual prefix. Essentially, the accuracy of reconstructed audio provides a constraint for AVMM to learn the latent visual-textual alignment. The reconstruct-based paradigm has another windfall, allowing AVMM to transfer to both multi-modal and uni-modal settings.

While the reconstruct-based paradigm implicitly learns the visual-textual alignment, we propose Dual Counterfactual Contrastive Learning (DCCL) to directly optimize the visual-audio alignment score and promote the robustness of reconstructed audio. We introduce distribution-based contrastive learning to further improve the accuracy of reconstructed audio and dependency-based contrastive learning with token-wise diversity-aware weights to provide modality-aware constraint from the perspective of text generation. Then, we apply the above module and a base audio-visual text generation network to the meta-learning framework, named TAVT, to empower AVMM with the ability to accrue knowledge across domains, which would assist in building internal multimodal representations broadly suitable for many domains. Our main contributions are as follows:

- We are the first one to study the transferable audio-visual text generation task.
- We introduce the audio-visual meta-mapping network that aligns domain-invariant low-level concepts between visual and a universal auditory semantic space.
- Experimental results on both cross-dataset and cross-category benchmarks demonstrate the effectiveness of our models.

2 Related Work

Audio-Visual Learning. In the past few years, there have been several works that focus on audio-visual learning. (Arandjelovic and Zisserman, 2018) used a two-stream neural network to find the most similar visual area to the current audio clip. Some works (Hu et al., 2019, 2020) employed contrastive learning to match the visual and audio components. Recently, (Liu et al., 2021; Cheng et al., 2023) proposed a framework for cross-modal representation learning with a discrete embedding space that was shared amongst different modalities and promoted model interpretability. The above approaches focus on the correlation of audio-visual pairs. While we aim to align the visual of different domains with a universal auditory semantic space and enhance multi-modal transfer learning.

Audio-Visual Text Generation. The mainstream audio-visual text generation task is video captioning, which has attracted many researchers (Venugopalan et al., 2015; Le et al., 2020; Ye et al., 2022; Jin et al., 2022a). (Hao et al., 2018) proposed multimodal feature fusion strategies to integrate audio information into models. (Guo et al., 2019; Iashin and Rahtu, 2020a) proposed an attention mechanism to combine the visual and audio features for better knowledge representation. (Tian et al., 2019) introduced an audio-visual controller to manipulate the parameters and generate diverse modality-aware captions. (Rahman et al., 2019) utilized the idea of cycle consistency to build a model with visual and audio inputs. (Iashin and Rahtu, 2020b) encoded the feature representation of audio and speech for a specific event proposal and produces a caption. Compared to our work, none of the above address the problem of domain shift and suffer severe degradation when deployed to a low-resource domain. Furthermore, they all take audio as supplementary information for visuals while we attempt to utilize the audio-visual correlation to minimize the domain discrepancy.

3 Methods

We aim to train a model that can learn and quickly adapt to new multimodal domains with limited labeled data under the meta-learning setting. In the next sections, we will first define the audio-visual meta-learning setting, then explain our architecture, and finally describe how it is used during training and inference time.

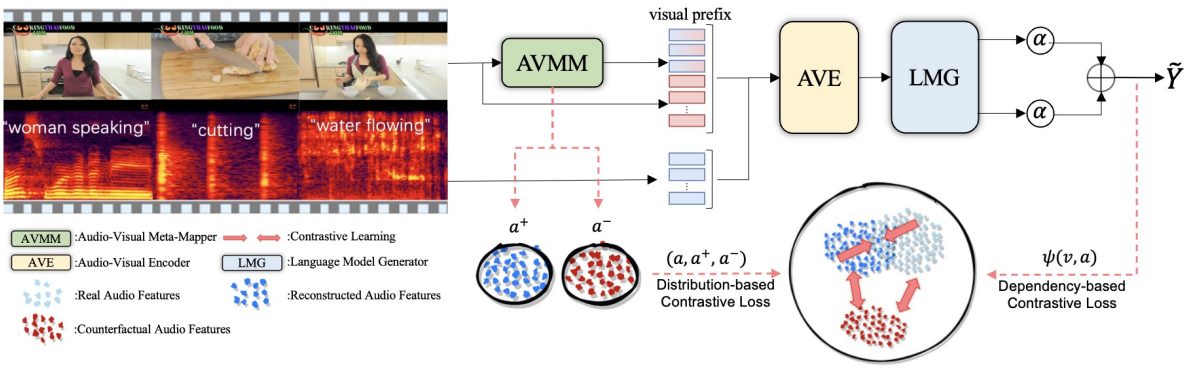

Figure 2: The architecture of our text generation network with counterfactual contrastive learning.

3.1 Problem Formulation

We define a meta-training stage, where the model is trained on the meta-train $D_{meta-train}$ partition and a separate meta-test stage where the performance is measured on the $D_{meta-test}$ partition. For the support set, there are k samples chosen for each of the $N-1$ randomly sampled domains, and for the query set, there are m samples from the remaining domain, with $m > k$. The set D_i is defined as $D_i = (v_1^i, a_1^i, t_1^i), ...(v_k^i, a_k^i, t_k^i)$, where v_j^i is the visual feature, a_j^i represents the audio features, and t_j^i is the output text, *i.e.,* a caption to the video.

3.2 Model Architecture

The architecture that we present in this paper is modular and consists of three components: the meta-mapper, an audio-visual encoder, and a language model as illustrated in Figure 2.

Audio-Visual Meta-Mapper Network. Intuitively speaking, low-level visual concepts in different domains often share similar sounds, *e.g.,* footsteps, laughing, and rain. Therefore, we propose an audio-visual meta-mapping network (AVMM) to map different visuals across domains into a universal auditory semantic space and as well as addressing shifts in the semantic distribution.

We first introduce the universal auditory semantic space which has audio clusters learned from Flickr(Thomee et al., 2015). In particular, we cut the audio into units with fixed time length T and run the clustering algorithm to find the k-centres of all audio features as the audio clusters. We define the audio clusters as $M = \{m_1, m_2, ..., m_k\}$. These audio clusters could serve as a bridge between the audio and visual and assist in learning quickly new domains by observing only limited labeled examples.

To map the visual features into the latent space

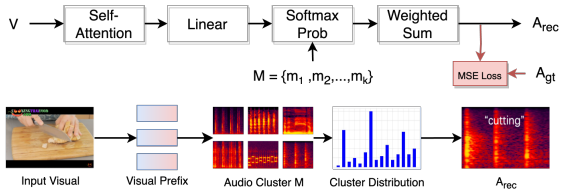

Figure 3: Illustration of Audio-Visual Meta-Mapper. Given a visual input, we map the visual to a visual prefix and combine clusters by weighting $m_1, m_2, ..., m_k$ to predict A_{rec}. The accuracy of A_{rec} provides a constraint for the self-attention layer to map an audio-aligned visual prefix.

of the audio clusters, inspired by prompt learning(Lester et al., 2021), we introduce a set of l learnable tokens $[p_1, p_2, ..., p_l]$. These tokens are called visual prefix for the audio clusters. Particularly, considering that the audio and visual are aligned in temporal, we also cut the frame sequence into l clips with time length T and obtain the embedding $[c_1, c_2, ..., c_l]$ by pooling the features of each clip. Then, we apply a mapper with self-attention (SA) on the c_i to obtain the visual prefix p_i as follows:

$$c_i = \frac{1}{|T|} \sum_{t \in T} v_t$$
$$p_i = \mathrm{SA}(c_i) \quad (1)$$

Then, we develop a reconstructor to reproduce the audio features from the visual prefix, and the accuracy of reconstructed audio provides a constraint for the self-attention layer to retrieve cross-domain invariance from the visual features c_i, and accumulate it into p_i. To achieve this, the reconstructor learns to combine audio clusters from M to reconstruct dynamic audio features $A_{rec} = \{a_{rec}^1, a_{rec}^2, ..., a_{rec}^l\}$. The predicted audio features a_{rec}^i is a weighted sum of audio clusters $m_1, m_2, ..., m_k$, where the weights are predicted

by applying a linear layer ϕ to p_i followed by a softmax function to normalize weights, as follows:

$$a_{rec}^{i} = \sum_{k=1}^{K} w_k m_k, \quad \text{where}, w_k = \text{SoftMax}(\phi_k(p_i)) \tag{2}$$

During meta-training, we apply the mean square error (MSE) loss $\mathcal{L}_{rec}$ on the reconstructed audio A_{rec} against the ground-truth audio A to update the parameters in the meta-mapper network and shared across all domains in $D_{meta-train}$.

Audio-Visual Encoder. Formally, given a sequence of video frames, we first extract a sequence of the frame features $V = \{v_1, v_2, \ldots, v_m\}$ and the audio features $A = \{a_1, a_2, \ldots, a_m\}$. Then, we prepend this visual prefix to the frame features, yielding the following sequence $V^{'} = [p_1, ..., p_l, v_1, ..., v_m]$. After that, we adopt a self-attention module to learn visual representation $f_v^{'} \in \mathbb{R}^d$ and audio representation $f_a^{'} \in \mathbb{R}^d$ as follows: $f^{'} = \text{MHA}(f, f, f)$, where, $f \in \{V, A\}$ and $\text{MHA}(\cdot)$ denotes multi-head attention (Vaswani et al., 2017). Then we apply audio-visual cross-attention to identify attention across two different kinds of feature fields. For simplicity, we formulate this stage as:

$$x_t = \text{AV–Encoder}(f_i, f_j, f_j), \text{where}, i, j \in \{v, a\} \tag{3}$$

where $x_t \in \{x_{av}, x_{va}\}$. Illustratively, the details of the encoder are provided in Appendix A.

Language Model Generator. As opposed to the original Transformer's decoder, we introduced an α to evaluate the contribution of different modalities (audio and visual) to each word. At time step t, α_t is computed by measuring the relevance between the cross-attention of each modality and the previous words $Y = \{y_1, y_2, ..., y_{t-1}\}$ as follows:

$$\alpha_t = \sigma\left(W_t\left[Y, \text{MHA}\left(Y, x_t, x_t\right)\right] + b_t\right) \tag{4}$$

where $[\cdot, \cdot]$ indicates concatenation, σ is the sigmoid activation and W_t is a $2d \times d$ weight matrix. The decoder outputs caption $\tilde{Y}$ is defined as:

$$\tilde{Y} = \alpha_{av} \cdot \text{MHA}(Y, x_{av}, x_{av}) + \alpha_{va} \cdot \text{MHA}(Y, x_{va}, x_{va}) \tag{5}$$

With α_t, the model can provide interpretability for the audio-visual fusion strategy.

3.3 Counterfactual Contrastive Learning

Although the reconstruction-based paradigm provides a constraint for AVMM, it cannot directly optimize the visual-audio alignment scores. Therefore, we propose a Dual Counterfactual Contrastive Learning (DCLL) which constructs fine-grained supervision signals from counterfactual results to directly optimize the visual-textual alignment without relying on the quality of randomly-selected negative samples.

Distribution-based Contrastive Learning. Concretely, we take the reconstructed audio cues A_{rec} as positive samples A^+ and inverse the audio clusters M and weight matrix w_k pairings to construct counterfactual audio cues as negative samples A^-. Then, we illustrate the contrastive learning method with the causal triplet (A, A^+, A^-). Intuitively, we construct distribution-based contrastive learning as follows

$$\mathcal{L}_{dis} = -\log\left(\frac{e^{\left(s\left(A, A^+\right)/\tau\right)}}{\sum_{i=1}^{n} e^{(s(A, A_i)/\tau)}}\right) \tag{6}$$

where $s\left(p, q\right) = p^T q / \left\|p\right\| \left\|q\right\|$ denotes the dot product between l_2 normalized p and q; τ is the temperature parameter. The distribution-based contrastive learning further improves the accuracy and robustness of reconstructed audio.

Dependency-based Contrastive Learning. For the audio-visual text generation task, there exists a modality imbalance in natural language tokens as different tokens depend on different modalities and the reconstructed audio should also show similar dependence for different tokens. We consider the dependency-based contrastive loss to maintain consistency in the distribution of scores for the original and positive samples. First, the (A, A^+, A^-) paired with the V are fed into the audio-visual encoder to generate the joint embeddings of them. Then, we compute the dependence score $\psi\left(V, A\right) = \alpha_{av}/\alpha_{va}$ of original sample as the anchor r, the score $\psi\left(V, A^+\right)$ of the factual sample as the positive r^+ and the score $\psi\left(V, A^-\right)$ of the counterfactual sample as the negative r^-. Concretely, the contrastive loss is formulated as follows:

$$\mathcal{L}_{dep} = -\log\left(\frac{e^{\left(s\left(r, r^+\right)/\tau\right)}}{\sum_{i=1}^{n} e^{(s(r, r_i)/\tau)}}\right) \tag{7}$$

The dependency-based contrastive learning considers the different performances of audio-visual in text generation to further provide cross-modal constraints on the mapper.

Token-Wise Modality-Aware Weights. In order to identify some vague concepts like talking and

singing, we devise the token-wise modality-aware weights to encourage the model to use the corresponding modality in the text generation process. We obtain the association of each word with audio modality and visual modality as follows:

$$W_{ma}^{i} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\alpha_{av}^{i}}{\alpha_{va}^{i}}\right) \tag{8}$$

Where W_{ma}^{i} indicates the i-th word's weights and N is the data sample size. We apply the weight W_{ma} on the cross-entropy loss, as follows:

$$\mathcal{L}_{cap} = -\sum_{t=1}^{n} W_{ma} * \log(w_t = y_t | y_{1:t-1}) \tag{9}$$

where $log(w_t = y_t|y_{1:t-1})$ denotes the probability of predicting word w_t given the previously generated $y_{1:t-1}$.

3.4 Meta-Training and Inference

The holistic training procedure is shown in Algorithm 1. Here, for simplicity, we assume that our full model is defined as a function f_θ, which receives the visual features v and audio features a as input and produces y as output. The loss function, optimized per domain during training, is as follow:

$$\mathcal{L} = \mathcal{L}_{cap} + \mathcal{L}_{rec} + \lambda\mathcal{L}_{dis} + \mu\mathcal{L}_{dep}, \tag{10}$$

where the hyper-parameter λ and μ seek a trade-off between the two counterfactual-contrastive learning losses (details about the hyper-parameter can be found in the supplementary materials).

Meta-Training To meta-train the model, we randomly select $K-1$ specific domains in D as support set D_s, and the remaining domain as a query set D_q. When adapting to a new domain D_i, the trainable meta parameters θ become *domain-specific* parameters, namely θ_i. These domain-specific parameters are computed with N gradient-step updates, similar as in MAML(Finn et al., 2017), with the following rule for one gradient update: $\theta_i^{'} = \theta - \alpha\nabla_\theta\mathcal{L}_{D_i}(f_\theta)$. This is referred as the inner-loop update, where α is the hyperparameter for the step size. Next, the model meta-parameters θ are optimized for the performance of $f_{\theta_i^{'}}$, using the query set D_q samples and the domain-specific parameters $\theta_i^{'}$ as initialization of the model:

$$\min\sum\mathcal{L}_{D_i}(f_{\theta_i^{'}}) = \sum\mathcal{L}_{D_i}(f_{\theta-\alpha\nabla_\theta\mathcal{L}_{D_i}(f_\theta)}) \tag{11}$$

which is called outer-loop optimization. The meta-optimization across all domains D_i is performed using stochastic gradient descent update rule, as follows: $\theta \leftarrow \theta - \beta\nabla_\theta\sum\mathcal{L}_{D_i}(f_{\theta_i^{'}})$, where β is the step size hyperparameter.

Algorithm 1: Transferable Audio-Visual Text Generation

Input: k source domains $D=\{D_1, ..., D_k\}$
Output: Model parameters θ

while *not done* **do**
  Randomly select $(k-1)D_s \sim D$, and the remaining domain as D_q ;
  Sample $B_i \sim D_s$;
  foreach B_i **do**
    $\theta_i^{'} = \theta - \alpha\nabla_\theta\mathcal{L}_{B_i}(f_\theta)$
  end
  Sample $B_q \sim D_q$;
  $\theta_q^{'} = \theta - \alpha\nabla_\theta\mathcal{L}_{B_q}(f_\theta)$;
  Meta-optimization;
  $\theta = \theta - \beta\nabla_\theta\sum_{i=1}^{K-1}(\mathcal{L}_{D_i}(f_{\theta_i^{'}}) + \mathcal{L}_{D_q}(f_{\theta_q^{'}}))$
end
Meta-Test;
if *audio in* D_{target} **then**
  $\theta = \theta - \alpha\nabla_\theta\mathcal{L}(f_\theta)$;
else
  /* Frozen the AVMM */
  $\theta = \theta - \alpha\nabla_\theta\mathcal{L}_{cap}(f_\theta)$;
end

Meta-Test In the meta-test stage, we consider a new domain D_{target}, which also has a support set D_s for fast adaptation by fine-tuning the model meta-parameters θ to a given task, and a query set D_q to evaluate the model on this domain. Note that in audio-absent datasets like MSVD, we can freeze the parameters of the audio-visual meta-mapper network and utilize the reconstructed audio features to boost the performance.

4 Experiment

4.1 Datasets and Metrics.

Datasets For transferable audio-visual text generation, we design two benchmark datasets. A cross-domain benchmark was constructed based on MSR-VTT(Xu et al., 2016) containing 20 categories as well as multimodal audio and video streams. We divided a new 10-domain dataset based on MSR-VTT categories as shown in Table 1 (more detailed information about the dataset can be found in Appendix B). For cross-dataset benchmark, we use MSVD(Chen and Dolan, 2011)

Domain	$D_{meta-train}$	Domain	$D_{meta-test}$
News	1727	Animation	816
Movie	1652	Music	733
Sports	1623	Animal	613
Cooking	985	Kids	558
Traffic	815	Beauty	478

Table 1: The number of videos for 10 domains reorganized from MSRVTT.

which consists of 1,970 video clips collected from YouTube, and MSR-VTT† which consists of the whole $D_{meta-test}$ as target domains.

Metrics. We evaluate the methods across all four commonly used metrics for video captioning: BLEU-n (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (ROUGE, 2004), CIDEr (Vedantam et al., 2015). We follow the standard practice to use the Microsoft COCO evaluation server (Chen et al., 2015).

4.2 Train Details.

Feature Extraction. For the visual feature, we use the ResNet-101 (He et al., 2016) model pre-trained on ImageNet as the backbone feature extractor for frames. The frames are used as the input of the CNN without re-sizing or cropping. The audio features are extracted by the VGGish(Hershey et al., 2017). For the sentences, to simplify the implementation, we use a joint vocabulary containing words in both the source domain and the target domain. And words appearing less than 3 times are replaced with a special token.

Proposed Method Settings. The hidden size is 1024 for all the multi-head attention mechanisms. The numbers of heads and attention blocks are 8. For meta-training, we adopt Adam with a fixed inner learning rate of 0.0001 and outer learning rate of 0.001. We train the source domains with a batch size of 32. For the meta-test, we use the beam-search method with a beam width of 5 to generate the predicted sentences during testing. We train TAVT on an NVIDIA GeForce RTX 2080, for TAVT each epoch takes around 4 hours.

4.3 Performance Evaluation

Compared Models. To the best of our knowledge, there is no work investigating transferable audio-visual text generation. So we first choose the state-of-the-art video caption approaches based on two different approaches: (1)The RNN-based models: RecNet (Wang et al., 2018), AVAF(Guo et al., 2019), MARN(Pei et al., 2019), AVIC(Tian et al., 2019), SGN(Ryu et al., 2021) and SHAN(Deng et al., 2022). (2) The Transformer-based models Att-TVT(Chen et al., 2018) and SBAT (Jin et al., 2020). Then, for a fair comparison, these methods are all trained on $D_{meta-train}$ **with the same meta-learning framework as TAVT** and tested on the target domain.

Evaluation Results. In Table 2 we report the performance of our method in comparison with the aforementioned competitors on the cross-datasets benchmark. As it can be observed: (1) Our method outperforms all compared methods on all metrics by a large margin on MSR-VTT† and MSVD. (2) In particular, AVIC and Att-TVT focus on designing complex multimodal fusion strategies to learn visual-audio representations, but leave the audio invariance unexploited. TAVT uses audio as a supervisory signal to align visual information in different domains, focusing on transferring the invariant in audio to visual prefix. Therefore, our model outperforms them by a significant margin (+1.5%$\sim$9.6% of CIDEr on MSR-VTT†). (3) Note that in the MSVD which has only a visual stream, we freeze the parameters of AVMM and use the reconstructed audio instead the real audio *i.e.,* TAVT(A_{rec}). TAVT still performs well (+13.5% of CIDEr on MSVD), which indicates that the meta-mapper network has accumulated domain-sharing knowledge through meta-training. In other words, the frozen meta-mapper network can produce discriminative visual prefix and reconstruct informative audio features even in the absence of real audio supervision.

In Table 3 we also report the performance on cross-categories benchmark. TAVT outperforms all other methods on five domains, which indicates that our method generalizes well. In particular, for some low-resource domains with only a few labeled data such as kids and beauty, other methods suffer from severe performance degradation, while TAVT outperforms them by a large margin (+3.2% on kids and +4.5% on beauty).

4.4 Ablation Studies

To analyze the effect of different components, we conduct ablation studies on the MSR-VTT† dataset. The following variants of our method are evaluated:

Effectiveness of audio-visual meta-mapper. To evaluate the advantage of AVMM and audio features. We first remove both AVMM and other sub-

Methods	→MSR-VTT† (mutli-modal)					→MSVD (uni-modal)				
	BLEU-1	BLEU-4	M	R	C	BLEU-1	BLEU-4	M	R	C
RecNet(Wang et al., 2018)	71.1	40.8	26.9	59.5	43.4	79.8	52.9	34.4	70.6	83.3
AVAF(Guo et al., 2019)	72.2	40.7	27.3	60.0	44.9	-	-	-	-	-
MARN(Pei et al., 2019)	73.7	40.2	27.9	60.2	47.6	83.6	50.1	35.7	72.5	93.7
AVIC(Tian et al., 2019)	74.0	41.4	28.1	61.2	48.6	-	-	-	-	-
SGN(Ryu et al., 2021)	75.8	41.2	28.1	60.6	50.2	84.4	53.3	35.6	72.6	95.2
SHAN(Deng et al., 2022)	76.2	40.9	28.2	60.1	51.5	84.2	53.5	35.4	72.8	95.7
Att-TVT(Chen et al., 2018)	74.9	40.3	28.1	60.3	48.5	-	-	-	-	-
SBAT(Jin et al., 2020)	75.4	40.9	28.3	60.9	50.4	82.0	53.6	35.5	72.2	91.1
TAVT (A_{rec})	78.3	41.8	28.3	61.8	52.6	**84.7**	**53.9**	**36.1**	**73.3**	**96.8**
TAVT	**78.5**	**42.1**	**28.6**	**61.9**	**53.0**	-	-	-	-	-

Table 2: Performance comparisons of two transfer tasks on the cross-datasets benchmark. All methods use the $D_{meta-train}$ as the source domain and transfer to →MSVD and →MSR-VTT†. The best results are bold.

Domain	Methods	B@4	M	R	C
→Animation	SBAT	44.7	29.2	64.2	47.6
	SHAN	45.5	29.6	64.7	48.6
	TAVT	**47.3**	**30.6**	**65.8**	**50.4**
→Music	SBAT	42.7	27.9	61.2	43.8
	SHAN	43.8	28.6	62.4	44.5
	TAVT	**45.5**	**29.2**	**63.1**	**46.1**
→Animal	SBAT	36.6	26.0	56.1	46.7
	SHAN	37.8	26.4	56.8	47.7
	TAVT	**38.3**	**27.0**	**58.1**	**48.4**
→Kids	SBAT	40.5	22.2	57.6	44.4
	SHAN	41.1	23.9	58.0	45.6
	TAVT	**42.7**	**26.1**	**60.4**	**47.6**
→Beauty	SBAT	31.6	24.1	52.9	27.3
	SHAN	33.1	24.5	53.5	28.5
	TAVT	**35.0**	**25.8**	**55.0**	**31.8**

Table 3: Performance comparisons of five transfer tasks on the cross-categories benchmark. The complete experimental results are shown in Appendix C.2.

Method	B@1	B@4	M	R	C
w/o. audio	75.3	37.9	26.3	59.7	47.2
w/o. AVMM	77.1	39.6	27.5	60.0	50.4
w/o. real audio	78.0	41.8	28.2	61.4	52.4
TAVT	**78.5**	**42.1**	**28.6**	**61.9**	**53.0**

Table 4: Ablation studies about audio features.

Methods	B@1	B@4	M	R	C
w/o. MAML	76.4	39.8	26.8	59.8	50.2
w/o. $\mathcal{L}_{rec}$	76.9	40.8	27.2	61.8	51.6
w/o. $\mathcal{L}_{dis}$	77.5	41.4	28.0	61.4	52.2
w/o. $\mathcal{L}_{dep}$	77.3	41.6	28.5	61.9	52.4
w/o. ma	77.8	41.8	28.4	61.5	52.7
TAVT	**78.5**	**42.1**	**28.6**	**61.9**	**53.0**

Table 5: Ablation studies about different modules.

modules which are related to audio and only use the visual information as the lower bound (**w/o. audio**). Then, we give the results without AVMM (**w/o. AVMM**). We also report the result that retains AVMM but uses the reconstructed audio instead of real audio (**w/o. real audio**).

Table 4 shows that while audio contains information that is complementary to visual and can improve performance somewhat (**w/o. audio** vs **w/o. meta-mapper**), the more important for transfer learning is the cross-domain invariance which contained in audio and provided the supervised signal to align visual in different domains (**w/o. meta-mapper** vs TAVT). In addition, the reconstructed audio can ultimately exhibit an upper bound on performance close to that achieved using real audio, demonstrating the accuracy and validity of the meta mapper network (**w/o. real audio** vs TAVT).

Effectiveness of different modules. The results in Table 5 illustrate that the constraints provided by the accuracy of the reconstructed audio features are effective and critical (**w/o.** $\mathcal{L}_{rec}$). The counterfactual thinking is helpful and can further improve the accuracy by +0.8% on CIDEr (**w/o.** $\mathcal{L}_{dis}$). Moreover, optimizing the cross-modal relationship between audio and visual from a text generation perspective can further improve the performance of the model (**w/o.** $\mathcal{L}_{dep}$). In addition, **w/o. MAML** denotes the model without meta-learning. We can observe that TAVT performs significantly better than **w/o. MAML**.

Effectiveness of token-wise modality-aware weight. To investigate how token-wise modality-aware weight improves the performance from the perspective of linguistics, we visualize some token's weight and the text generation process in Appendix E. The result in Table 5 shows that token-wise modality-aware weights can further improve the performance of TAVT by optimizing the association of text and audio-visual modalities.

Figure 4: Illustrations of text generated by SHAN and TAVT. The last one on MSVD has only a visual stream. The red word highlights the advantages of TAVT.

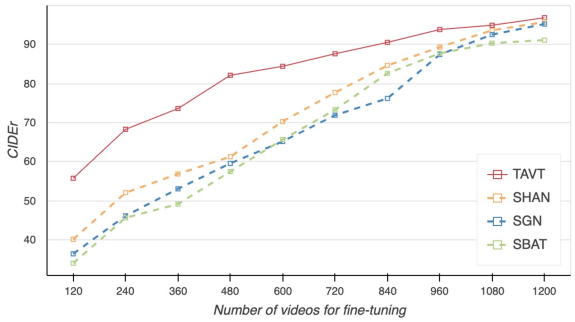

Figure 5: The performance of TAVT and previous methods under different sizes of labeled videos. As the results show, TAVT can achieve competitive results with only 40% of the data.

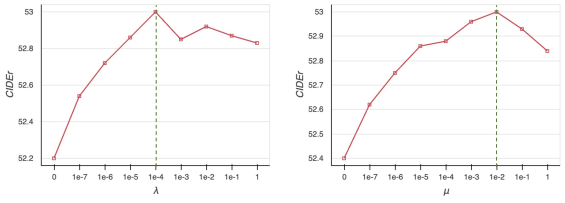

Figure 6: Model performances under different λ and μ hyper-parameters on MSR-VTT†

The performance in low-resource domain. To quantitatively verify the performance in the low resource domain, we compare TAVT and previous methods under different sizes of labeled video as shown in Figure 5. We observe that TAVT consistently outperforms the other methods and can achieve close to full performance with only 40% of the training data, while other methods require about 70% of the training data.

Hyper-Parameter Analysis. To seek the trade-off between the DCCL, we introduce the hyper-parameter λ and μ. Figure 6 shows that the model achieves the best performance when λ=1e-4 and μ=1e-2, suggesting that proper hyper-parameters are crucial to achieve good performance.

4.5 Qualitative Analysis

To qualitatively verify the effectiveness of our TAVT, we display the results of TAVT in the multi-modal and uni-modal settings. As shown in Figure 4, TAVT can accurately describe low-level concepts such as "chirping", "rainy" and "piano" with the help of audio-enhance visual prefix. When transferred to a uni-modal domain where missing audio, the AVMM can capture the correlation audio-visual correlation and reconstruct audio features that are semantically relevant to the visual.

5 Conclusion

In this paper, we first study the task of transferable audio-visual text generation tasks. To mitigate multimodal content domain shift, we observe that low-level visual concepts have similar sounds in different domains and propose a novel framework TAVT with two technical contributions. The first one is the audio-visual meta-mapper, which can transfer the domain-invariant concept information within the universal auditory semantic space into the visual prefix. Moreover, the well-trained audio-visual meta-mapper can also reconstruct semantic audio features in audio-absent mode. We then apply dual counterfactual contrastive learning to directly optimize the visual-audio alignment. Extensive experiments on both cross-datasets and cross-domains benchmarks verify the effectiveness of our model. Furthermore, our TAVT framework can be transferred to other text generation tasks *e.g.*, video QA in a plug-and-play fashion.

6 Limitation

We identify a few limitations of the current work. Our approach still suffers from biases in the training data and may produce incorrect output or lead to an inaccurate understanding of multi-modal content. And a large-scale audio-visual pre-trained model is a promising direction toward more advanced and cheaper approaches for transfer learning, which we leave for future study.

7 Ethics Statement

We adopt the widely-used datasets that were produced by previous researchers and followed all relevant legal and ethical guidelines for their acquisition and use. Besides, we recognize the potential influence of our technique. When deployed our approach will have to record, store and process video and audio information related to human activities, which will have privacy implications for some application domains. We are committed to conducting our research ethically and ensuring that our research is beneficial. We hope our work can inspire more investigations for transfer learning on multi-modal tasks and wish our framework can serve as a solid baseline for further research.

8 Acknowledge

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, and Yiwise.

References

Relja Arandjelovic and Andrew Zisserman. 2018. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Ming Chen, Yingming Li, Zhongfei Zhang, and Siyu Huang. 2018. Tvt: Two-view transformer network for video captioning. In *Asian Conference on Machine Learning*, pages 847–862. PMLR.

Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE international conference on computer vision*, pages 521–530.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

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*.

Jincan Deng, Liang Li, Beichen Zhang, Shuhui Wang, Zhengjun Zha, and Qingming Huang. 2022. Syntax-guided hierarchical attention network for video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):880–892.

Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. 2022. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7212–7222.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Ningning Guo, Huaping Liu, and Linhua Jiang. 2019. Attention-based visual-audio fusion for video caption generation. In *2019 IEEE 4th ICARM*, pages 839–844. IEEE.

Wangli Hao, Zhaoxiang Zhang, and He Guan. 2018. Integrating both visual and audio cues for enhanced video caption. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *2017 icassp*, pages 131–135. IEEE.

Chiori Hori, Takaaki Hori, and Jonathan Le Roux. 2021. Optimizing latency for online video captioningusing audio-visual transformers. *arXiv preprint arXiv:2108.02147*.

Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257.

Di Hu, Zheng Wang, Haoyi Xiong, Dong Wang, Feiping Nie, and Dejing Dou. 2020. Curriculum audiovisual learning. *arXiv preprint arXiv:2001.09414*.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989.

Vladimir Iashin and Esa Rahtu. 2020a. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*.

Vladimir Iashin and Esa Rahtu. 2020b. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959.

Tao Jin, Siyu Huang, Ming Chen, Yingming Li, and Zhongfei Zhang. 2020. Sbat: Video captioning with sparse boundary-aware transformer. *arXiv preprint arXiv:2007.11888*.

Tao Jin, Zhou Zhao, Peng Wang, Jun Yu, and Fei Wu. 2022a. Interaction augmented transformer with decoupled decoding for video captioning. *Neurocomputing*, 492:496–507.

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.

Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. 2020. Hierarchical multimodal attention for end-to-end audio-visual scene-aware dialogue response generation. *Computer Speech & Language*, 63:101095.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Alexander H Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2021. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438*.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. 2019. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8347–8356.

Tanzila Rahman, Bicheng Xu, and Leonid Sigal. 2019. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8908–8917.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of Association for Computational Linguistics, Spain*.

Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. 2018. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814.

Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D Yoo. 2021. Semantic grouping network for video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2514–2522.

Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2021. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447*.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 1(8).

Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. 2019. Audio-visual interpretable and controllable video captioning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*.

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.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *ICCV*, pages 4534–4542.

Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7622–7631.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael Lyu. 2022. Understanding and improving sequence-to-sequence pretraining for neural machine translation. *arXiv preprint arXiv:2203.08442*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5288–5296.

Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. 2022. Hierarchical modular network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17939–17948.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5119.

Appendix

This appendix contains four sections. (1) Appendix A introduce the detail about audio-visual encoder. (2) Appendix B introduces the details of the benchmark construction. (3) Appendix C provides the complete performance comparison on cross-categories benchmark (3) Appendix D discusses the design of universal auditory semantic space (Appendix D.1) and visualize the universal Auditory semantic space (Appendix D.2) (4) Appendix E provide the analysis of token-wise modality-aware weights and visualize the text generation process.

A Encoder

For the visual input $V^{'} = [p_1, ..., p_l, v_1, ..., v_m]$ and audio input $A = [a_1, ..., a_m]$, we first adapt the self-attention layer SelfAtt to learn the visual representation and audio representation as follows:

$$f_v^{'} = \text{SelfAtt}(V^{'}, V^{'}, V^{'}) \quad (12)$$
$$f_a^{'} = \text{SelfAtt}(A^{'}, A^{'}, A^{'}) \quad (13)$$

The SelfAtt layer contains multi-head attention which can be calculated by multiple single heads:

$$\text{MHA}(F, F, F) = \text{Concat}(h_1, h_2, ..., hh), \quad (14)$$
$$h_i = \text{ATT}(FW_i^Q, FW_i^K, FW_i^V) \quad (15)$$

where, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\frac{d}{h}}$ and $W^1 \in \mathbb{R}^{d \times d}$. h_i denotes the i-th head and h is the number of heads. ATT represents scaled dot-product attention as $\text{ATT}(Q, K, V) = \text{softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}})$. Then we apply audio-visual cross-attention to identify attention across two different kinds of feature fields as follows:

$$x_{av} = \text{CrossAtt}(f_v^{'}, f_a^{'}, f_a^{'}) \quad (16)$$
$$x_{va} = \text{CrossAtt}(f_a^{'}, f_v^{'}, f_v^{'}) \quad (17)$$

where CrossAtt layer is similar to SelfAtt contains multi-head attention.

B Details about Datasets

To more adequately validate the effectiveness of our proposed approach, we first regrouped the 20 categories of MSR-VTT and reorganization a new 10-domain dataset based on the MSR-VTT (Xu et al., 2016) categories information as shown in Table 6. The number of videos regrouped 10 domains is shown in Figure 7. We use five of them as source domains and the remaining five domains ("Animation", "Music", "Animal", "Kids" and "Beauty") as target domains. And a random sampling strategy is applied in the target domain for dataset split.

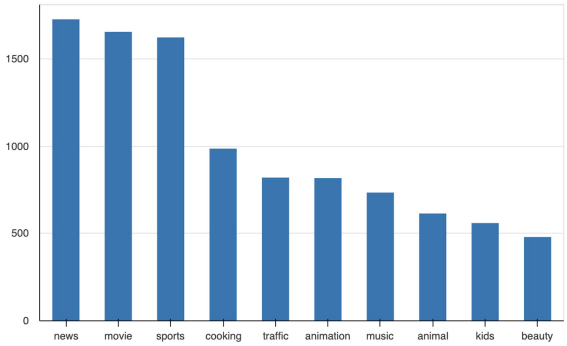

Figure 7: The distribution of video numbers in our regrouped 10 domains.

Domain	Category
News	howto, education, science, news
Movie	tv shows, movie, doc, ads
Sports	people, sports, travel
Cooking	food, cooking
Traffic	vehicle
Animation	gaming, animation
Music	music
Animal	animal
Kids	kids
Beauty	beauty

Table 6: Statistics of our 10 regroup captioning domains based on the MSR-VTT dataset.

C Experiment Results

C.1 Baseline Setting

To the best of our knowledge, there is no work investigating transferable audio-visual text generation, thus we choose the state-of-the-art video caption approach and use MAML on top of it to construct the baseline. We compare the performance of the TAVT with that of state-of-the-art methods based on two different approaches: (1)The RNN-based models: RecNet (Wang et al., 2018) which adds a reconstructor to reconstruct the visual features from the generated caption, AVAF(Guo et al., 2019) combined different multimodal fusion methods and attention mechanism, MARN(Pei et al., 2019) which equips with a memory consisting of words and corresponding visual contexts, AVIC(Tian et al., 2019) introduce audio-visual controller to balance the importance between audio and

Methods	→Animation					→Music				
	BLEU-1	BLEU-4	M	R	C	BLEU-1	BLEU-4	M	R	C
RecNet(Wang et al., 2018)	72.7	42.8	27.4	62.0	45.3	71.6	41.5	26.5	59.1	40.1
MARN(Pei et al., 2019)	75.5	44.5	28.3	63.4	47.2	74.6	43.4	27.6	60.3	42.6
AVIC(Tian et al., 2019)	75.2	44.8	28.7	63.6	47.1	74.9	43.0	27.7	60.5	42.9
SGN(Ryu et al., 2021)	75.6	45.1	29.2	64.2	48.5	75.8	43.4	28.0	61.9	44.5
SHAN(Deng et al., 2022)	76.2	45.5	29.6	64.7	48.6	76.0	43.6	28.2	61.8	44.2
Att-TVT(Chen et al., 2018)	72.4	44.3	29.0	63.8	46.3	73.1	42.2	27.8	61.7	42.0
SBAT(Jin et al., 2020)	72.1	44.7	29.2	64.2	47.6	74.8	42.7	27.9	61.2	43.8
TAVT	**76.8**	**47.3**	**30.6**	**65.8**	**50.4**	**79.0**	**45.5**	**29.2**	**63.1**	**46.1**

Table 7: The results of performance comparisons. All methods use the $D_{meta-train}$ as the source domain and transfer to vehicle and music domains. The best results are bold.

Methods	→Animal					→Kids				
	BLEU-1	BLEU-4	M	R	C	BLEU-1	BLEU-4	M	R	C
RecNet(Wang et al., 2018)	72.9	35.6	24.3	53.5	43.6	73.0	37.2	21.4	55.0	41.8
MARN(Pei et al., 2019)	74.7	36.9	25.7	55.1	45.0	75.5	38.7	22.5	55.8	43.3
AVIC(Tian et al., 2019)	75.0	36.8	25.9	55.7	45.6	74.9	39.1	22.8	56.3	43.2
SGN(Ryu et al., 2021)	76.6	38.1	26.6	56.2	47.3	76.1	40.7	23.3	57.6	45.1
SHAN(Deng et al., 2022)	76.3	37.8	26.4	56.8	47.7	76.5	41.1	23.9	58.0	45.6
Att-TVT(Chen et al., 2018)	74.7	36.0	25.8	56.3	46.2	75.3	40.2	21.7	57.1	44.3
SBAT(Jin et al., 2020)	75.8	36.6	26.0	56.1	46.7	75.1	40.5	22.2	57.6	44.4
TAVT	**78.5**	**38.3**	**27.0**	**58.1**	**48.4**	**77.3**	**42.7**	**26.1**	**60.4**	**47.6**

Table 8: The results of performance comparisons. All methods use the $D_{meta-train}$ as source domain and transfer to animal and kids domains. The best results are bold.

Methods	→Beauty				
	BLEU-1	BLEU-4	M	R	C
RecNet	60.7	28.9	21.8	51.3	24.6
MARN	64.4	30.5	23.0	52.2	26.4
AVIC	64.8	30.7	23.5	52.0	26.3
SGN	65.3	33.3	24.5	53.7	28.6
SHAN	65.0	33.1	24.5	53.5	28.5
Att-TVT	62.5	31.1	23.7	52.3	26.8
SBAT	63.4	31.6	24.1	52.9	27.3
TAVT	**67.8**	**35.0**	**25.8**	**55.0**	**31.8**

Table 9: The results of performance comparisons. All methods use the $D_{meta-train}$ as source domain and transfer to animal and kids domains. The best results are bold.

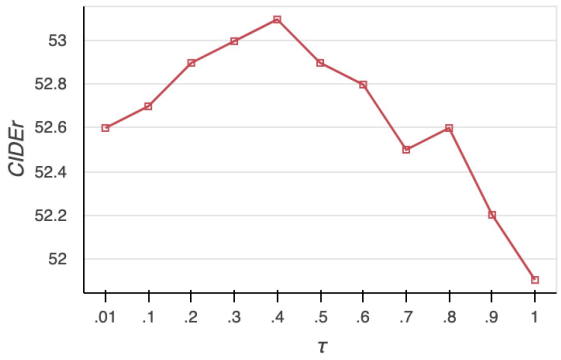

Figure 8: Impact of Temperature τ on MSRVTT†.

C.2 Performance on Cross-Categories Benchmark

visual modalities, SGN(Ryu et al., 2021) which encode a video into semantic groups and SHAN(Deng et al., 2022) which use syntax-guided hierarchical attention to integrate visual and sentence-context features. (2) The Transformer-based models Att-TVT(Chen et al., 2018) which fuses the modalities from video and text with attention mechanism, SBAT (Jin et al., 2020) which uses boundary-aware pooling operation to reduce the redundancy.

To further validate the generalizability of the proposed method, we also conduct experiments on five reorganized $D_{meta-test}$ domains. From the results in Table 7,8 and 9, we observe that our method outperforms other methods in all five target domains, which indicates that our method has good generalization and can compensate for the performance degradation caused by limited label data in low-resource domains such as kids and beauty.

Cluster Number	B@1	B@4	M	R	C
10	77.3	41.4	28.2	61.2	52.0
100	78.5	42.1	28.6	61.9	53.0
200	78.2	41.7	28.5	61.7	52.3
300	77.0	41.2	27.9	61.1	51.8

Table 10: The Comparison of different cluster number.

C.3 Hyper-Parameter Analysis

We evaluate 11 different values τ from 0.01 to 1.0 on MSRVTT† and report the results in Figure 8. It shows that the performance achieves the best when τ is set to 0.4 and becomes poor when τ is too small or too large. This result suggests that a proper τ value is crucial to achieving good performance.

D Analysis of Universal Auditory Semantic Space

D.1 Construction Details

Audio Clusters. We design our model by introducing a universal auditory semantic space consisting of a set of audio clusters. We are interested in natural environmental sounds. We download videos from videos on Flickr(Thomee et al., 2015) and extract their sounds. We downloaded over 750,000 videos from Flickr, which provides over a year (377 days) of continuous audio. The only pre-processing we do on the sound is to extract the spectrogram from the video files and subtract the mean. We extract spectrograms for approximately five seconds of audio and obtain the audio clip features by VGGish and select conv4_1 as the extraction layer.

Selection of Clustering Algorithm. To obtain the audio cluster, we apply K-means (MacQueen et al., 1967) to the extracted audio features. Specifically, K-Means require a manual setting of cluster number values. Thus, we experimented with different cluster numbers. From the results shown in Table 10, we can see that when setting the cluster number as 100 achieve better results on our extracted audio features from Flickr.

D.2 Visualization of Audio Clusters.

In Figure 9, we visualize the semantic space. We can see that there are a large number of low-level concepts shared across different domains. The sounds of these low-level concepts are similar but have significant visual differences. This provides a guarantee for the effectiveness of our approach.

E Analysis of Token-Wise Modality-Aware Weights

Modality Imbalance. There exists a modality imbalance in natural language tokens as different tokens depend on different modalities. For example, some nouns of objects like "shirts", "lamps" and "flowers" are only visually related, but some objects like "guitar" and "alarm clock" can make sounds that are auditory related. Moreover, the verbs which indicate the low-level concepts like "talking", "hit" and "kick" are obviously auditory related. And we visualize the word clouds in Figure 10.

Visualization of Text generation. To investigate how the token-wise modality aware weight improves the performance from the perspective of linguistics, we visualize the text generation process. In Figure 11, we present the modality dependency scores of each word. Firstly, we can find that words have different dependencies on different modalities in the process of generation, which proves the existence of modal imbalance. For example, some quantifiers (a, group) and some nouns (children) rely more on the visual modality, while guitar and sing rely more on the audio modality. There are also some prepositions and conjunctions that often rely on the context text for a generation. Secondly, comparing **w/o. ma** and TAVT, we can find that token-wise modality-aware weights can optimize the dependence of different words on different modalities, encouraging the model to use the correct modality to generate words, which can help identify some vague concepts such as talking and singing.

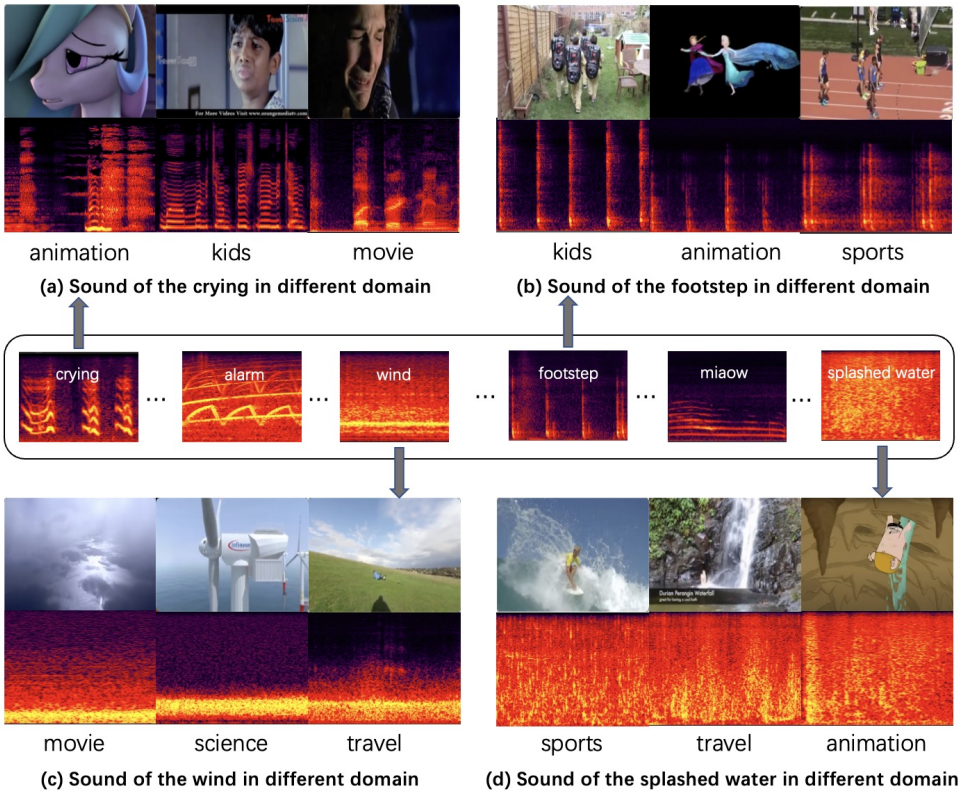

Figure 9: Universal Auditory Semantic Space Visualization: We visualize a few clusters from the semantic space and how they align to different domains. On top of the spectrograms (yellow/red heatmaps), we also show the original video frame related to the sound.

(a) Word clouds for words related to audio

(b) Word clouds for words related to visual.

Figure 10: Word clouds for audio-visual words where font size indicates the audio-visual relevancy of words.

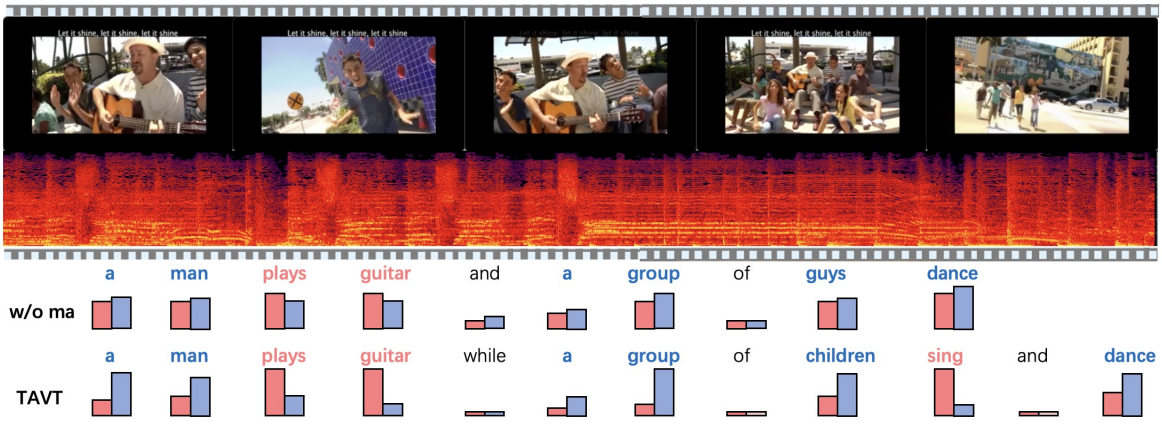

Figure 11: The visualization of the dependency of different words on audio and visual modalities during text generation. Red indicates audio modality and blue indicates visual modality. The Higher bars represent higher dependency scores.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

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

☑ A3. Do the abstract and introduction summarize the paper's main claims?
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?

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?
Section 4.2

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?
Section 4.2

☐ 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?
Not applicable. Left blank.

☑ 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.)?
Section 4

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.