Towards Multimodal Video Paragraph Captioning Models Robust to Missing Modality

Anonymous ARR submission

Abstract

Video paragraph captioning (VPC) involves generating detailed narratives for long videos, utilizing supportive modalities such as speech and event boundaries. However, the existing models are constrained by the assumption of constant availability of a single auxiliary modality, which is impractical given the diversity and unpredictable nature of real-world scenarios. To this end, we propose a Missing-Resistant framework MR-VPC that effectively harnesses all available auxiliary inputs and maintains resilience even in the absence of certain modalities. Under this framework, we propose the Multimodal VPC (MVPC) architecture integrating video, speech, and event boundary inputs in a unified manner to process various auxiliary inputs. Moreover, to fortify the model against incomplete data, we introduce *DropAM*, a data augmentation strategy that randomly omits auxiliary inputs, paired with *DistillAM*, a regularization target that distills knowledge from teacher models trained on modality-complete data, enabling efficient learning in modality-deficient environments. Through exhaustive experimentation on YouCook2 and ActivityNet Captions, MR-VPC has proven to deliver superior performance on modality-complete and modality-missing test data. This work highlights the significance of developing resilient VPC models and paves the way for more adaptive, robust multimodal video understanding.

1 Introduction

Video Paragraph Captioning (VPC) (Park et al., 2019) is a fundamental video-language understanding task that requires the model to generate paragraph-level captions for minutes-long videos. Besides raw video frames, there exist several auxiliary modalities that can potentially serve as supplementary inputs, such as speech inputs utilized in Vid2Seq (Yang et al., 2023b), flow features used in MART (Lei et al., 2020), and event boundaries (the start and end timestamps of the events) leveraged in various models (Zhou et al., 2018b; Yamazaki et al., 2022a,b, etc). Despite the growing performance of these models, we notice that they assume to have access to the same auxiliary modality during both training and testing, which contradicts reality. In real-world scenarios, the availability of modalities undergoes dynamic changes, which leads to the following two issues for the models developed under the unrealistic assumption.

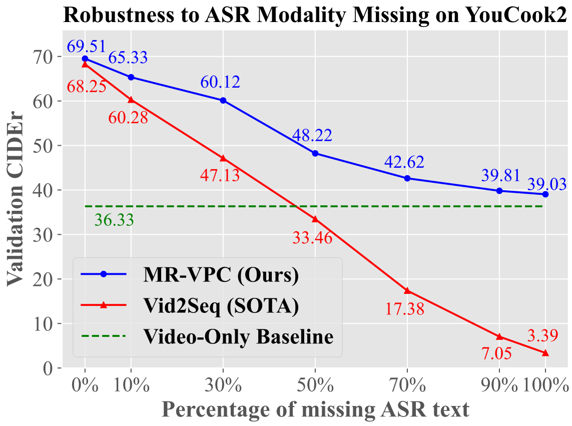


Figure 1: The performance of the previous SOTA model Vid2Seq drastically declines as the percentage of ASR text missing grows. In contrast, our **MR-VPC** consistently achieves superior performance in both modality-complete and modality-missing environments.

Issue-1: Under-utilization of available modalities. Since a specific auxiliary modality is solely considered during training, the models fail to leverage unseen modalities that may emerge at test time. For example, VLCap and VLTinT (Yamazaki et al., 2022a,b) cannot employ transcribed speech, which is proven extremely beneficial in Vid2Seq (Yang et al., 2023b); conversely, Vid2Seq cannot make use of event boundaries, which contain rich information about the temporal structure of videos. **Issue-2: Vulnerability to missing modality in noisy environments.** The performance of these models may degrade drastically when the required auxiliary modality is absent or of low quality, which is common in real-world situations. For instance, Liu and Wan (2021) find that the VPC models relying on event boundaries yield significantly lower performance when the ground-truth event boundaries are replaced with learned ones. Besides, we observe that the state-of-the-art model Vid2Seq is vulnerable to the missing of automatically transcribed speech (ASR texts) as depicted in Figure 1.

In response to **issue-1**, we design a multimodal VPC (**MVPC**) architecture to integrate the inputs from multiple modalities. Concretely, **MVPC** first encodes the two auxiliary modalities (*i.e.*, tokenized event boundaries and transcribed speech) into a unified textual feature space using a shared text encoder. Then, the textual features are fused with the video features before entering the language decoder to generate paragraph captions. Further, to alleviate **issue-2**, we devise two training strategies to enhance the robustness of our model to missing modalities. Firstly, we simulate the absence of auxiliary modalities by randomly dropping the inputs (named *DropAM*) during training. This approach reduces the model's reliance on auxiliary inputs and improves generalization in noisy situations. Second, to take full advantage of the auxiliary modalities, we propose to perform multimodal knowledge distillation (Hinton et al., 2015) (referred to as *DistillAM*) where the model trained on modality-complete data acting as the teacher and the model operating in modality-missing situations learning as the student. By combining **MVPC**, *DropAM* and *DistillAM*, we present a **M**ultimodal noise-**R**esistant **V**ideo **P**aragraph **C**aptioning framework (**MR-VPC**).

Experimental results on two benchmarks demonstrate the superiority of **MR-VPC** in handling both modality-complete and modality-incomplete data. Notably, **MR-VPC** is tailored for the challenging VPC task and substantially outperforms prior robustness-oriented methods studied for classification tasks. To our knowledge, this work pioneers formulating VPC as a multimodal learning problem with noisy inputs and presents practical solutions that enable VPC systems to utilize inputs from diverse modalities while remaining robust even when parts of them are missing.

2 Related Work

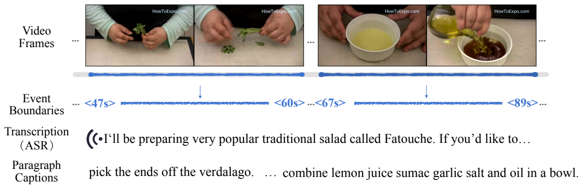
Video Paragraph Captioning (VPC) VPC is a widely studied video-language understanding task involving producing paragraph-level captions for long videos lasting for minutes (Park et al., 2019). Existing VPC models commonly incorporate additional auxiliary information alongside video frames as inputs, such as transcribed speech (Yang et al., 2023b) and event boundaries (Zhou et al., 2018b; Yamazaki et al., 2022a,b, etc). Liu and Wan (2021) and Song et al. (2021) build VPC models for raw videos without event boundaries, but their models still underperform those utilizing auxiliary modalities. To the best of our knowledge, our work takes the first step to utilize both transcribed speech and event boundaries for VPC in an end-to-end manner, and we are the first to study the robustness of VPC models to noisy inputs with missing modalities.

Robustness to Missing Modality As multimodal neural networks are vulnerable to missing modality (Ma et al., 2022), recent years have seen a surge of studies on enhancing model robustness on modality-incomplete data across various multimodal tasks (Woo et al., 2022; Lee et al., 2023; Wei et al., 2023; Yuan et al., 2023, etc). In terms of methodology, researchers have explored approaches such as modality fusion strategy search (Ma et al., 2022), data augmentation in the form of modality dropout (McKinzie et al., 2023), and regularization objectives (Woo et al., 2022; McKinzie et al., 2023). However, existing efforts are limited to relatively simple classification tasks, and model robustness to missing modality in more complex language generation tasks like VPC is yet to be explored. We have found that simply applying the existing approaches in other tasks does not achieve satisfactory results in VPC and bridge this research gap by developing training strategies customized for VPC in our **MR-VPC** framework, which will be discussed in § 3 and § 4.

3 Methodology

3.1 Problem Formulation

An instance in a VPC dataset can be formulated as (V_i, A_i, E_i, C_i), where V, A, E, C stand for video frames, ASR texts, event boundaries, and the caption, respectively. An example from the YouCook2 (Zhou et al., 2018a) dataset is illustrated in Figure 2. We assume that the video modality V is always available at test time and the auxiliary modalities A and E are likely to be affected by noise in the wild. Given N_A and N_E as the noise functions for A and E (*e.g.*, random missing in the context of our study on missing modality), respectively, for a model $F(V, A, E)$ trained on the clean



I'll be preparing very popular traditional salad called Fatouche. If you'd like to…

pick the ends off the verdalago. … combine lemon juice sumac garlic salt and oil in a bowl.

Figure 2: The composition of an instance in the multi-modal VPC task from the validation set of YouCook2.

training set $D_{\text{tr}} = \{(V_i, A_i, E_i, C_i), 1 \le i \le n_{\text{tr}}\}$, where n_{tr} is the size of the training data, our target is to maximize the performance on the noisy test set $D_{\text{te}} = \{(V_i, N_A(A_i), N_E(E_i), C_i), 1 \le i \le n_{\text{te}}\}$, where n_{te} is the size of the test data.

3.2 MVPC Model Framework

Overview Overall, as illustrated in Figure 3, our multimodal video paragraph captioning (MVPC) model consists of four modules: the video encoder E_v to encode V, the text encoder E_t to encode the concatenation of A and E, a fusion module E_f that merges visual and textual features, and a text decoder D_t that generates the caption C.

Video Encoder The video encoder E_v encodes the video sequence of F frames $x_v \in \mathbb{R}^{F \times H \times W \times C}$, where H, W and C are the height, width, and the number of channels, respectively, and outputs the video embedding sequence $E_v(x_v) \in \mathbb{R}^{F \times d}$, where d is the embedding size. Concretely, we use a CLIP ViT-L/14 (Radford et al., 2021) image encoder to encode each frame and then feed the frame features into a 12-layer Transformer (Vaswani et al., 2017) for temporal interaction.

Text Encoder To resolve **issue-1**, we expect the model to be capable of modeling both A and E inputs end to end. Thus before feeding A and E into the text encoder E_t, we adopt the relative time tokenization (Yang et al., 2023b) to map continuous timestamps into discrete time tokens denoting the percentage progress. Then E_t transforms the concatenation of the ASR sequence and event boundary sequence x_t consisting of n tokens in total into the text embedding sequence $E_t(x_t) \in \mathbb{R}^{n \times d}$.

Fusion Module and Text Decoder At the end of the workflow, the text decoder D_t generates the target caption sequence in an auto-regressive manner, conditioned on the encoder embeddings produced by the fusion module E_f merging $E_v(x_v)$ and $E_t(x_t)$. Specifically, for E_f, we adopt a parameter-free concatenation operation; for E_t and D_t, we employ the T5v1.1-base encoder-decoder model (Raffel et al., 2020).

Test Modalities	YouCook2		ActivityNet	
	METEOR	**CIDEr**	**METEOR**	**CIDEr**
V+E+A	23.11	74.13	14.09	42.29
V+A	21.05 (-2.06)	59.55 (-14.58)	12.24 (-1.85)	29.71 (-12.58)
V+E	12.46 (-10.65)	8.77 (-65.36)	12.91 (-1.18)	43.14 (+0.85)
V	6.79 (-16.32)	3.42 (-70.71)	11.64 (-2.45)	26.08 (-16.21)

Table 1: The performance of the vanilla MVPC model on YouCook2 and ActicityNet Captions in different modality missing settings.

Weight Initialization To benefit from large-scale pretraining, we initialize the model with the Vid2Seq weight pretrained on YT-Temporal-1B (Zellers et al., 2022) [1]. Note that our work differs from Vid2Seq in terms of the task context and research goal. We aim at the VPC task that generates textual paragraph-level captions C from the input modalities V, A and E, where A and E are likely to be missing, while Vid2Seq is originally designed for the dense video captioning task where the inputs are V and A (without considering missing modality) and the outputs are C and E. To establish a baseline for comparison, we re-implement Vid2Seq and fine-tune its pretrained weights for the VPC task (details in Appendix B). This allows us to evaluate the performance improvement achieved by our proposed framework. Note that MVPC is **not** a simple extension of Vid2Seq, as our general framework to incorporate A and E unitedly is agnostic to the underlying structure and applies to other vision-language foundation models.

3.3 Training Strategies of MR-VPC

As the vanilla training of MVPC does not consider potential noise in the inference stage, it suffers from severe performance drops facing missing modality (**issue-2**), as shown in Table 1. For instance, the absence of A results in a 65.36 (88.17% relatively) CIDEr drop on YouCook2; the missing of E causes a 12.58 (29.75% relatively) CIDEr decline on AcitivityNet. [2] In light of this weakness, we explore the following training strategies to enhance the model's resilience to missing modality (the model trained with them is referred to as **MR-VPC** later).

3.3.1 *DropAM*: Drop Auxiliary Modalities

Since the missing modality can be viewed as a distribution shift from the training data, a fundamental

[1] Available at this link.

[2] We find that the ASR data of ActivityNet contains little useful cues and show small negative effects, so we nullify the ASR input of ActivityNet at test time later.

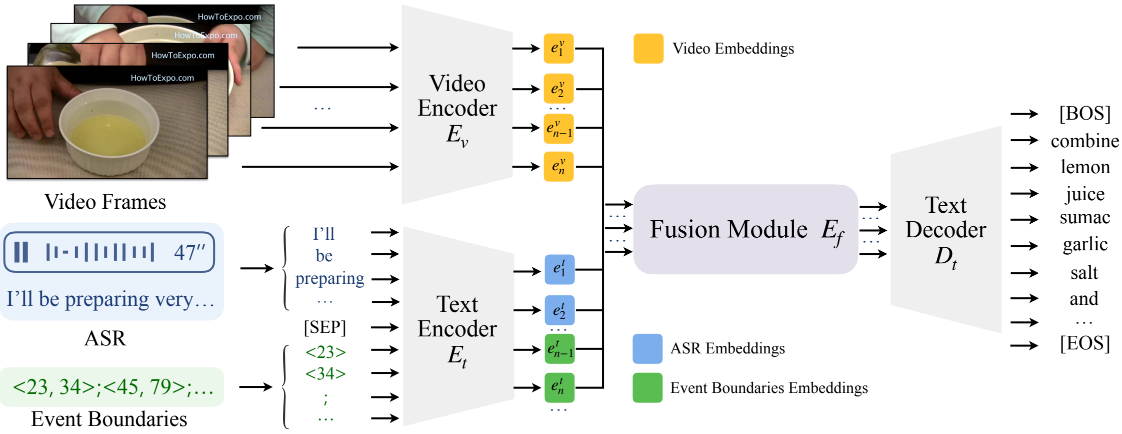


Figure 3: The overview diagram of our MVPC (multimodal video paragraph captioning) framework.

idea to enhance model robustness is simulating the noise during training. To this end, we randomly drop the auxiliary modalities A and E to reduce the dependence of the model on them. Specifically, we transform the original training set D_{tr} to $\hat{D}_{\text{tr}} = \left\{\left(V_i, \hat{N}_A\left(A_i\right), \hat{N}_E\left(E_i\right), C_i\right), 1 \le i \le n_{\text{train}}\right\}$, in which $\hat{N}_A$ and $\hat{N}_E$ are the proxy noise functions that random replace A_i and E_i with a default null character at probabilities p_A and p_E, respectively:

$$\hat{N}_A\left(A_i\right) = \begin{cases} '', & p \le p_A \\ A_i, & p > p_A \end{cases}, \hat{N}_E\left(E_i\right) = \begin{cases} '', & p \le p_E \\ E_i, & p > p_E \end{cases}, \tag{1}$$

where p is a random variable uniformly drawn from the range $[0, 1]$. We use $p_A = p_E = 0.5$ as the value works generally well in practice. Please see the discussion about their effects in Appendix D.

3.3.2 *DistillAM*: Learning from the Teacher with Modality-Complete Data

Solely applying *DropAM* turns the model training into a multitask learning process involving subtasks with different input conditions, which possibly adds to the learning difficulty and compromises the performance on modality-complete data. Therefore, we resort to knowledge distillation (Hinton et al., 2015), a learning paradigm that transfers the knowledge from teacher models with better conditions, such as more training data and a larger number of parameters, to student models without these advantages. In our problem, we consider the vanilla MVPC model trained on the modality-complete training set D_{tr} as the teacher model F_t, and our goal is to transfer the knowledge learned by F_t to the **MR-VPC** model that likely faces missing modality as the student model F_s. In early trials, we have found that distilling from word-level logits (WordKD) achieves limited performance gains in our task. Therefore, inspired by the sequence-level knowledge distillation (SeqKD) (Kim and Rush, 2016) studied in machine translation, we create a new training set D_{kd} by replacing the ground-truth caption C with the predictions given by F_t based on the modality-complete data:

$$D_{\text{kd}} = \left\{\left(V_i, A_i, E_i, F_t\left(V_i, A_i, E_i\right)\right), 1 \le i \le n_{\text{tr}}\right\}, \tag{2}$$

and then construct the augmented training set $D_{\text{aug}} = D_{\text{tr}} \bigcup D_{\text{kd}}$ by merging D_{kd} and the original training data D_{tr}. It is notable that this procedure named *DistillAM* is orthogonal to the noise simulation process *DropAM* in § 3.3.1, so they can be applied together, *i.e.*, the random noise can be injected into the augmented training data D_{aug} in the training phase in the way stated in § 3.3.1.

3.3.3 Connection to Prior Strategies for Multimodal Classification Tasks

Although MASD (McKinzie et al., 2023), the state-of-the-art approach to enhance model robustness to missing modality in classification problems, also takes the form of modality dropout and knowledge distillation, it differs from our solutions in essence. Concretely, MASD performs self-distillation, namely aligning the predicted probabilities on modality-complete and modalities-incomplete data output by the same model under training. In contrast, we use a fixed teacher model trained on modality-complete data, which facilitates the efficient learning of the student model in the challenging VPC task. We will show the advantage of our MR-VPC over MASD and its variant MASD+Wise-FT (McKinzie et al., 2023) in § 4.2.2.

4 Experiments

4.1 Experimental Setup

Evaluation Protocol Following Yang et al. (2023b), we use CIDEr (C) (Vedantam et al., 2015) and METEOR (M) (Banerjee and Lavie, 2005) metrics to evaluate the accuracy of generated captions. For measuring diversity, we use 4-gram repetition (R@4) (Xiong et al., 2018) following Liu and Wan (2021) and Yamazaki et al. (2022a,b). Besides these metrics based on n-gram matching commonly used in previous works, we also report advanced model-based metrics in § 5.1.

Benchmarks We conduct main experiments on YouCook2 (Zhou et al., 2018a) and ActivityNet Captions (Krishna et al., 2017), two widely studied VPC benchmarks containing paragraph-level captions and annotated event boundaries. We report the evaluation metrics on the validation set of YouCook2 and the *as-test* split of ActivityNet Captions (see Appendix A for details).

Acquisition of ASR Data For ActivityNet Captions, we adopt the ASR data provided by Iashin and Rahtu (2020) from the YouTube ASR system. For YouCook2, we obtain the ASR data using the *whisper-timestamped* tool (Louradour, 2023) based on Whisper (Radford et al., 2022) (the *small.en* model with 244M parameters) and dynamic time warping (Giorgino, 2009).

Model Training and Inference We train the model for 40 epochs on YouCook2 and 20 epochs on ActivityNet Captions using a batch size of 32. The model is trained with the Adam (Kingma and Ba, 2015) optimizer to minimize cross-entropy loss with an initial learning rate of 2e-4 with cosine annealing. For training efficiency, we freeze the image encoder in our experiments unless otherwise mentioned, so the number of trainable parameters is 314M. The weight decay is 5e-2 and we clip the maximum norm of the gradient to 1.0. We uniformly sample 100 frames at resolution 224×224 pixels for the video input and the ASR text sequence is truncated at the max length of 1000. Temporally consistent random spatial augmentation (Qian et al., 2021) is applied. The inference beam search size is 4 and the repetition penalty is 1.2. See more details in Appendix B.

Evaluation Settings We mainly report results in three representative test settings: (1) **the modality-complete setting** where the auxiliary modalities A and E are not affected by any noise; (2) **the video-only setting** where both A and E are missing, which is a harsh but realistic setting (in the real world, most users do not enter the video's event boundaries E; A is also possibly missing, *e.g.*, when the ASR system does not support the conversation language); (3) **the random-missing setting** where A and E are both randomly missing at the probability of 50% independently.

Model	Training Strategies		Test Modalities			
	DropAM	*DistillAM*	V+E+A	V+E	V+A	V
MVPC	✗	✗	**74.13/23.11**	8.77/12.46	59.55/21.05	3.42/6.79
-	✓	✗	60.40/22.67	35.17/16.94	64.87/22.54	36.73/16.53
MR-VPC	✓	✓	69.51/22.83	**39.03/16.97**	**69.37/22.59**	**38.37/16.86**

Table 2: The effect of our training strategies with different available modalities at test time on the YouCook2 dataset. CIDEr / METEOR metrics are reported.

Baselines We compare our models with a wide array of baselines and categorize them according to the input modalities in their original settings:

- **V:** The Vid2Seq model finetuned on only the video modality, named Vid2Seq (V); Soft-NMS (Bodla et al., 2017), ESGN (Mun et al., 2019), Memory Transformer (Song et al., 2021), and VPCSum (Liu and Wan, 2021); MART, MART$^{\text{COOT}}$, Vanilla Transformer, and Transformer-XL. The last four models use event boundaries generated by ESGN at test time as done in Liu and Wan (2021).
- **V+E:** VLTinT (Yamazaki et al., 2022b), VLCap (Yamazaki et al., 2022a), MART (Lei et al., 2020), MART$^{\text{COOT}}$ (Ging et al., 2020), Vanilla Transformer (Zhou et al., 2018b), and Transformer-XL (Dai et al., 2019).
- **V+A:** Vid2Seq (Yang et al., 2023b).

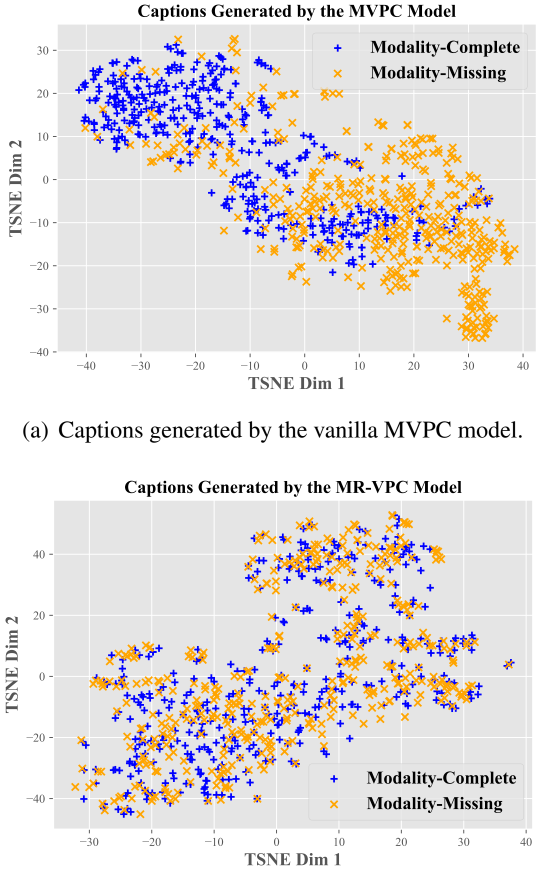
4.2 Results and Analysis

4.2.1 Comparing MVPC and MR-VPC

Our training strategies remarkably boost the model's robustness to missing modality while maintaining the performance in the modality-complete setting. Before comparing our model with baselines, we first examine the effectiveness of our training strategies described in § 3.3. As the results displayed in Table 2, the vanilla MVPC model without these training strategies is extremely susceptible to missing modality at test time, but the MR-VPC model equipped with these techniques shows substantially improved robustness to missing modality with only minimal performance sacrifice on the modality-complete test data. For instance, MVPC disastrously fails in the video-only setting (the CIDEr falls to 3.42), while MR-VPC yields a CIDEr value of 38.37. We also affirm the validity of each strategy by comparing MR-VPC with the model trained with only the *DropAM* strategy (the last two rows of Table 2). As shown, although *DropAM* boosts the

BERTScore↑	YouCook2	ActivityNet
MVPC	82.37	91.72
MR-VPC	**91.30**	**94.16**

Table 3: The average BERTScore similarities between captions generated in modality-complete and video-only test scenarios.



(a) Captions generated by the vanilla MVPC model.

(b) Captions generated by the MR-VPC model.

Figure 4: Visualization of the SimCSE embeddings of the captions generated under modality-complete and modality-missing (video-only) scenarios.

Model	YouCook2			ActivityNet		
	C ↑	M ↑	R@4 ↓	C ↑	M ↑	R@4 ↓
MVPC (*Ours*)	**74.13**	**23.11**	0.82	**43.14**	13.91	0.67
MR-VPC (*Ours*)	69.51	22.83	**0.57**	41.01	13.84	**0.51**
Baselines						
Vid2Seq	68.25	23.01	0.75	30.77	12.51	0.82
Vid2Seq (V)	36.33	16.79	0.79	28.87	12.38	0.57
VLTinT	48.70	17.94	4.29	31.13	**17.97**	4.75
VLCap	49.41	17.95	5.16	30.29	17.48	4.18
MART	35.74	15.90	4.39	22.16	15.57	5.44
MART$^{\text{COOT}}$	46.06	18.17	6.30	28.19	15.99	6.64
Vanilla Trans.	38.00	11.55	-	21.33	15.54	7.45
Memory Trans.	-	-	-	26.55	15.64	2.75
Trans.-XL	26.40	14.80	-	21.71	14.91	8.79
VPCSum	23.92	15.11	0.65	24.33	15.84	1.54

Table 4: Evaluation results under the modality-complete setting. ↑ indicates larger is better and ↓ indicates lower is better. The best result is highlighted in **bold**.

model robustness on modality-incomplete data, it significantly hurts the performance on modality-complete data (the CIDEr declines from 74.13 to 60.40); *DistillAM* not only further advances the robustness to missing modality, but also help preserve the performance in the modality-complete setting, as it raises the CIDEr metric to 69.51.

MR-VPC shows higher prediction consistency between modality-complete and modality-missing scenarios. To intuitively understand the impact of our training strategies, we compare the BERTScore (Zhang et al., 2019) similarities between the captions generated on modality-complete and video-only data by the vanilla MVPC and MR-VPC models. As listed in Table 3, MR-VPC exhibits substantially higher similarity scores, which indicates that it is capable of generating more consistent predictions, regardless of the availability of auxiliary modalities. Furthermore, we visualize the SimCSE embeddings (Gao et al., 2021) [3] of the generated captions on YouCook2 using t-SNE (Van der Maaten and Hinton, 2008) in Figure 4, where we observe that the captions generated by MVPC form two distinct clusters depending on whether modality-missing occurs, but those produced by MR-VPC appear in pairs and seem hard to distinguish based on the test scenario. The visualization further proves that *DropAM* and *DistillAM* contribute to the consistency of the predictions.

4.2.2 Comparison with Advanced Systems

Our MVPC and MR-VPC obtain superior performance in the modality-complete setting. We present the evaluation results in the modality-complete setting in Table 4 and observe that our models markedly advance the state-of-the-art on most metrics. In terms of captioning accuracy, we elevate the CIDEr metric from 68.25 (Vid2Seq) to 74.13 on YouCook2 and from 31.13 (VLTinT) to 43.14 on ActivityNet; regarding diversity, we achieve the lowest R@4 repetition scores below 1.0. These results support the necessity to fully leverage the auxiliary modalities A and E (**issue-1**) and the effectiveness of our MVPC model frameowork. We notice that VLTinT and some earlier baselines do better in terms of METEOR on AcitivyNet than Vid2Seq and our models, but we contend that ours and Vid2Seq are better models for two reasons: (1) CIDEr is a more reasonable metric because it accounts for the importance of different n-grams

[3] We use the unsup-simcse-roberta-large model.

Model	YouCook2			ActivityNet		
	C ↑	M ↑	R@4 ↓	C ↑	M ↑	R@4 ↓
MVPC (*Ours*)	3.42	6.79	2.31	26.08	11.64	0.60
MR-VPC (*Ours*)	**38.37**	**16.86**	**0.57**	**31.37**	12.06	**0.58**
Baselines						
Vid2Seq	3.39	6.81	2.80	30.01	12.18	0.73
Vid2Seq (V)	36.33	16.79	0.79	28.87	12.38	**0.58**
Memory Trans.	-	-	-	26.55	15.64	2.75
VPCSum	23.92	15.11	0.65	24.33	**15.84**	1.54
SoftNMS	18.18	13.67	4.94	22.58	14.93	10.17
ESGN	21.85	15.74	6.51	17.01	13.37	4.94
Vanilla Trans.	20.95	15.11	7.04	16.88	13.37	2.85
Trans.XL	14.24	12.67	3.20	20.73	14.89	7.45
MART	16.56	13.44	4.63	20.16	14.94	6.09
COOT	19.67	14.21	5.99	21.83	14.67	1.54

Table 5: Evaluation results under the video-only setting.

Model	YouCook2			ActivityNet		
	C ↑	M ↑	R@4 ↓	C ↑	M ↑	R@4 ↓
MVPC (*Ours*)	33.31	15.70	1.55	33.55	12.86	0.59
MR-VPC (*Ours*)	**51.13**	**20.15**	0.74	**37.05**	13.01	**0.56**
Baselines						
Vid2Seq	33.46	14.19	1.46	29.93	12.48	0.75
Vid2Seq (V)	36.33	16.79	0.79	28.87	12.38	0.58
VPCSum	23.92	15.11	**0.65**	24.33	**15.84**	1.54

Table 6: Results under the random-missing setting.

and has shown higher consistency with human evaluation (Shi et al., 2022); (2) model-based metrics in § 5.1 and human study results in § 5.3 further corroborate the advantages of our models.

Our MR-VPC model performs significantly better in modality-missing settings than previous SOTA models. Given the figures displayed in Table 5 and Table 6, MR-VPC yields the best performance in the video-only and random-missing setting with substantial margins over baselines including those specially trained for the video-only setting such as Vid2Seq (V) (Yang et al., 2023b), VPCSum (Liu and Wan, 2021), and Memory Transformer (Song et al., 2021). This suggests that MR-VPC fulfills our objective of developing a robust VPC model capable of leveraging available auxiliary modalities while maintaining robustness even when they are missing in real-world scenarios.

Our MR-VPC shows the best cross-dataset generalization performance on the video-only Charades dataset. To further examine the cross-dataset generalization capability, we assess the models trained on ActivityNet Captions on the test set of the Charades (Sigurdsson et al., 2016), where only the video modality is available. As the results listed in Table 7, MR-VPC outperforms baselines in the zero-shot scenario where domain shift and missing modality occur simultaneously, further validating the strength of our approach.

Model	CIDEr	BERTScore	BARTScore
MVPC	6.79	87.08	-4.56
MR-VPC	**8.74**	**87.22**	**-4.47**
Vid2Seq	4.74	86.83	-4.62
Vid2Seq (V)	6.01	87.00	-4.48

Table 7: Zero-shot evaluation results on Charades (the model weights are trained on ActivityNet Captions).

Method	Test Modalities				Avg.
	V+E+A	V+E	V+A	V	
WordKD	64.50	30.62	65.33	27.21	46.92
MASD	67.95	32.98	68.72	33.47	50.78
MASD+WiSE-FT	68.90	34.96	**69.54**	32.54	51.49
MR-VPC (Ours)	**69.51**	**39.03**	69.37	**38.37**	**54.07**

Table 8: Comparison with other robustness-oriented methods with different available modalities at test time on YouCook2. CIDEr metrics are reported.

Our MR-VPC beats the SOTA robustness-oriented training methods in classification problems. As shown in Table 8, MR-VPC remarkably outperforms the state-of-the-art solutions towards robustness to missing modality in classification problems, *i.e.*, MASD and MASD+Wise-FT (McKinzie et al., 2023). This illustrates that our customized approaches for the VPC task make significant strides compared to simply incorporating existing techniques studied for other tasks previously. Besides, we observe that replacing the SeqKD with Word-KD leads to significant performance drops in all scenarios, which supports the rationality of using SeqKD in our *DistillAM* component.

4.3 Qualitative Results

Besides the above quantitative results, we provide qualitative evidence to support the superiority of our models. First, we find that MVPC and Vid2Seq tend to produce degenerated captions in the modality-missing setting, whereas the prediction of MR-VPC remains almost unchanged, as exemplified by the instance given in Table 14 in Appendix H. Moreover, even in the modality-complete setting, the Vid2Seq and VLTinT baselines often predict concepts that are not present in the video; in contrast, our MVPC and MR-VPC model produces fewer such hallucinations, as illustrated in Figure 5 in Appendix H.

5 Further Evaluation

5.1 Evaluation with Model-Based Metrics

Besides the n-gram-based metrics reported in § 4.2, we further compare our models with

Model	YouCook2			ActivityNet Captions				
	PPL ↓	BERT ↑	BART ↑	PPL ↓	BERT ↑	BART ↑	EMS ↑	EMS_{ref} ↑
VLTinT (Yamazaki et al., 2022b)	21.99	89.01	-3.91	30.97	88.03	-3.94	28.94	36.88
Vid2Seq (Yang et al., 2023b)	15.89	**90.58**	**-3.08**	24.68	88.71	-3.78	**29.54**	36.99
MVPC (Ours)	15.50	90.56	**-3.08**	18.77	**88.98**	**-3.56**	29.37	**37.21**
MR-VPC (Ours)	**15.11**	89.51	-3.49	**17.17**	88.85	-3.58	29.10	36.90

Table 9: The model-based metrics evaluated under the modality-complete setting. ↑ indicates higher is better and ↓ indicates lower is better. We highlight the best model in **bold**. We do not report EMScore on YouCook2 as the captions of YouCook2 are longer than the max length limit of CLIP, the backbone of the EMScore metric.

Noise Type	Low-Quality ASR			ASR Sentence Deletion			Event Deletion			Boundary Perturbation			Generated Boundary		
Metric	CIDEr	BERT	BART	CIDEr	BERT	BART	CIDEr	BERT	BART	CIDEr	BERT	BART	CIDEr	BERT	BART
Vid2Seq	60.39	90.35	-3.12	48.01	89.62	-3.31	68.25	90.58	-3.08	68.25	90.58	-3.08	**68.25**	90.58	-3.08
MVPC (Ours)	59.58	90.36	-3.13	48.95	89.66	-3.29	63.43	90.54	-3.11	**72.60**	90.57	-3.07	61.71	90.58	-3.07
MR-VPC (Ours)	**63.69**	**90.63**	**-3.08**	**53.59**	**90.04**	**-3.24**	**70.72**	**90.85**	**-3.02**	69.11	**90.86**	**-3.03**	67.02	**90.84**	**-3.03**

Table 10: The evaluation results under five forms of noise in auxiliary modalities.

Group1	MVPC	VLTinT	Equal
	56.0%	20.7%	23.3%
Group2	MR-VPC	VLTinT	Equal
	56.0%	18.7%	25.3%

Table 11: The average percentage of human preferences.

competitive baselines (Vid2Seq and VLTinT) using the following model-based metrics (details in Appendix C), as they align better with human preference (Shi et al., 2022): (1) **Perplexity (PPL)** for fluency; (2) **BERTScore** (Zhang et al., 2019) and **BARTScore** (Yuan et al., 2021) measuring prediction-reference similarity; (3) **EMScore** (Shi et al., 2022) for the matching extent of the prediction and the video frames and its extension $\textbf{EMS}_{\textbf{ref}}$. We present the results in Table 9 and find that our MVPC and MR-VPC obtain the best performance across most of these metrics. Notably, although VLTinT reaches the highest METEOR on ActivityNet, it falls behind our models and Vid2Seq on these metrics. We will further show the advantage of our models through human evaluation in § 5.3.

5.2 Generalization on Other Forms of Noise

Besides completely missing, the auxiliary modalities in the real world may also be affected by other weaker forms of noise, such as variations in ASR quality between the training and test phases. We further test our models and VidSeq under five types of noise: lower ASR quality and sentence deletion for A; event deletion, boundary perturbation, and generated boundaries for E (details in Appendix F). We present the results in Table 10 and see that although these forms of noise are not seen during training, our MR-VPC shows the best robustness in most cases, which again substantiates the generalizability of our training strategies. We believe that we will achieve even better robustness to these types of noises if we consider them in the choice of the proxy noise functions $\hat{N}_A$ and $\hat{N}_E$ in *DropAM*.

5.3 Human Evaluation

We conduct two groups of human evaluation, in which three annotators compare the captions generated by VLTinT and MVPC (or MR-VPC) in the modality-complete setting for 50 randomly sampled videos from the AcitivityNet Captions test set. They need to choose a caption showing higher consistency with the video content or mark that two captions are equally good (details in Appendix I). As shown in Table 11, our MVPC and MR-VPC significantly surpass VLTinT in pair-wise comparison, which again proves their superiority.

6 Conclusion

We present MR-VPC, a multimodal video paragraph captioning model capable of utilizing three input modalities (video, transcribed speech, and event boundaries) and keeping robust in the presence of missing modality. The MR-VPC framework comprises two key contributions: (1) the MVPC architecture, which seamlessly processes inputs from all three modalities in an end-to-end manner; (2) the incorporation of two training techniques, *DropAM* and *DistillAM*, which enhance the model's robustness when faced with missing modality. Through exhaustive experimental evaluation on YouCook2 and ActivityNet Captions datasets, we demonstrate the superiority of MR-VPC in various test scenarios, highlighting its practicality and efficacy in addressing the challenges of video paragraph captioning in real-world settings.

Limitations

We discuss the limitations of our work as follows. (1) Despite the outstanding performance of MR-VPC in modality-missing settings, it slightly lags behind our vanilla MVPC in the modality-complete setting. This is comprehensible because the optimization of the regularization targets introduced in *DropAM* and *DistillAM* may conflict with the learning on modality-complete data to some extent. We will conduct more explorations to reduce this gap. (2) We primarily study the absence (discussed in most of the main text) and other forms of noise (studied in § 5.2) in two main auxiliary modalities, namely transcribed speech and event boundaries, which do not cover all possible harsh test conditions in the wild. For future work, we intend to investigate the robustness of VPC models to other forms of data noise, such as video frame blurring, for a more comprehensive evaluation.

Ethics Statement

We believe that our proposal would contribute to the robustness and security of video captioning systems deployed in the open-world environment, as the absence and quality reduction of auxiliary modalities are common in practice. Our proposal also applies to other multimodal natural language generation tasks, *e.g.*, multimodal machine translation, on which we plan to conduct more studies in the future. Moreover, all pretrained models used in this work are publicly available, ensuring transparency and accessibility. Although we do not expect any direct negative consequences resulting from this paper, we hope to continue to build on our MR-VPC framework and develop stronger and safer multimodal VPC models in our future work.

References

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.

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. 2017. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618.

Toni Giorgino. 2009. Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(7).

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

J Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54.

Vladimir Iashin and Esa Rahtu. 2020. Multi-modal dense video captioning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 958–959.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal prompting with missing modalities for visual recognition. *arXiv preprint arXiv:2303.03369*.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2603–2614.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Hui Liu and Xiaojun Wan. 2021. Video paragraph captioning as a text summarization task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 55–60.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jérôme Louradour. 2023. whisper-timestamped. `https://github.com/linto-ai/whisper-timestamped`.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186.

Brandon McKinzie, Vaishaal Shankar, Joseph Yitan Cheng, Yinfei Yang, Jonathon Shlens, and Alexander T Toshev. 2023. Robustness in multimodal learning under train-test modality mismatch. In *International Conference on Machine Learning*, pages 24291–24303. PMLR.

Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. 2019. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6588–6597.

Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. 2019. Adversarial inference for multi-sentence video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17929–17938.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer.

Yuqing Song, Shizhe Chen, and Qin Jin. 2021. Towards diverse paragraph captioning for untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11245–11254.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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.

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.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857.

Shicai Wei, Chunbo Luo, and Yang Luo. 2023. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. 2022. Towards good practices for missing modality robust action recognition. *arXiv preprint arXiv:2211.13916*.

Yilei Xiong, Bo Dai, and Dahua Lin. 2018. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 468–483.

Kashu Yamazaki, Sang Truong, Khoa Vo, Michael Kidd, Chase Rainwater, Khoa Luu, and Ngan Le. 2022a. Vlcap: Vision-language with contrastive learning for coherent video paragraph captioning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3656–3661. IEEE.

Kashu Yamazaki, Khoa Vo, Sang Truong, Bhiksha Raj, and Ngan Le. 2022b. Vltint: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2211.15103*.

Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023a. Vidchapters-7m: Video chapters at scale. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023b. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2023. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748.

A Dataset Statistics

We conduct main experiments on YouCook2 (Zhou et al., 2018a) and ActivityNet Captions (Krishna et al., 2017). YouCook2 consists of 1,333 videos in the training set and 457 ones in the validation set. Each instance in YouCook2 has 7.7 event segments on average. ActivityNet Captions comprises 10,009 samples in the training set and 4,917 ones in the original validation set. Following the practice of Lei et al. (2020) and most of the baselines, we split the validation set to the *as-val* set of 2,460 videos and the *as-test* split of 2,457 videos. Each sample in ActicityNet Captions has 3.65 event segments on average. The average video length is 2.0 minutes in ActivityNet Captions and 5.3 minutes in YouCook2. Also, we test the cross-dataset performance on the test set of the Charades (Sigurdsson et al., 2016) dataset consisting of 1,760 videos. The average video length of Charades is 30s.

B More Implementation Details

Vid2Seq and Vid2Seq (V) We notice that the original Vid2Seq paper (Yang et al., 2023b) also reports the performance of Vid2Seq on the VPC task, but we have confirmed that the results are obtained by removing timestamp outputs from dense captioning outputs and they are inferior to the results we get by fine-tuning the Vid2Seq weight specifically on the VPC task where the inputs are V and A. Therefore, we report our fine-tuning results as the performance of Vid2Seq in the main text. Moreover, to get a competitive baseline in the video-only setting, we fine-tune the Vid2Seq pretrained weight in this setting and report the results as the performance of Vid2Seq (V) in the main text. The training schemes for Vid2Seq and Vid2Seq (V) follow the setup stated in § 4.1.

Our MVPC and MR-VPC During inference, we apply a length penalty of 1.0 for YouCook2 and ActivityNet Captions, and a length penalty of 0.6 for Charades. In the *DistillAM* strategy, when utilizing the MVPC model to generate training data for training the MR-VPC model, we keep the same inference hyperparameters. Notably, we notice that unfreezing top CLIP layers has minimal impact on the performance of MVPC and VidSeq in our preliminary experiments, but the choice significantly boosts the performance of MR-VPC. Thus, we unfreeze the last six CLIP layers in the video encoder when training MR-VPC models. In this situation, the total trainable parameters are 390M.

C Details of Model-Based Metrics

We use the following model-based automatic evaluation metrics:

- **Perplexity (PPL)**: To assess the fluency of the generated paragraph-level captions, we adopt the perplexity score produced by a pretrained language model gpt2-large (Radford et al., 2019) (774M parameters).

- **BERTScore** (Zhang et al., 2019) and **BARTScore** (Yuan et al., 2021) are two text generation metrics based on the similarities of BERT (Devlin et al., 2019) embeddings and the generation probabilities of the BART (Lewis et al., 2020) model, respectively. We use them for evaluating the consistency between generated captions and reference captions. Specifically, for BERTScore, we use the F1 score given by the roberta-large (Liu et al., 2019) pretrained model (335M parameters); for BART score, we use the facebook/bart-large-cnn model (406M parameters) trained on ParaBank2 (Hu et al., 2019). [4]

p_A	p_E	Test Modalities				Avg.
		V+E+A	V+E	V+A	V	
0.1	0.1	23.15	13.50	21.80	10.76	17.30
0.3	0.3	23.04	15.76	22.52	15.39	19.18
0.5	0.5	22.67	16.94	22.54	16.53	**19.67**
0.7	0.7	22.24	17.10	22.30	17.03	**19.67**
0.9	0.9	19.86	17.52	19.84	17.57	18.70

Table 12: The effect of the choice of drop rate p_A and p_E with different available modalities at test time on the YouCook2 dataset. Only the *DropAM* strategy is applied and METEOR metrics are reported.

- **EMScore** (Shi et al., 2022) is an automatic video captioning metric derived by matching the video frame embeddings and the text token embeddings produced by the CLIP (Radford et al., 2021) model. Besides the reference-free version EMScore (EMS for short), we also report the reference-based version EMS$_{\text{ref}}$ additionally considering the similarity of the prediction and the reference annotation. Concretely, we use the clip-vit-base-patch32 pretrained model (151M parameters) following Shi et al. (2022).

D Effect of Drop Rates p_A and p_E

Recall that p_A and p_E are the probabilities to be nullified for the ASR modality A and the event boundary modality E in our *DropAM* strategy in § 3.3.1. We enumerate the values of these two hyperparameters (called drop rates) in *DropAM* and report the CIDEr results on the validation set of YouCook2 in Table 12. We observe that large drop rates hamper performance in the modality-complete setting and small drop rates result in poor performance in the modality-incomplete setting. Generally, setting p_A and p_E around 0.5 strikes the balance relatively well and performs the best in terms of the average performance with different available modalities. Moreover, we have made similar observations on ActivityNet Captions in preliminary explorations. Therefore, we use p_A = p_E = 0.5 in our main experiments.

[4] Available at `https://github.com/neulab/BARTScore`.

Model	Test Modalities				Avg.
	V+E+A	V+E	V+A	V	
VidSeq-Concat	22.35	14.08	21.92	12.12	17.62
MR-VPC (Ours)	**22.83**	**16.97**	**22.59**	**16.86**	**19.81**

Table 13: Comparison with Vid2Seq-Concat (Yang et al., 2023a) on YouCook2. METEOR metrics are reported. When testing Vid2Seq-Concat without E, we trim the video into seven consecutive clips of the same length (seven is the average number of events in YouCook2).

E Comparison with Yang et al. (2023a)

Concurrent to our work, Yang et al. (2023a) extend Vid2Seq to incorporate both A and E for VPC (called *"video chapter generation given ground-truth boundaries"* in their paper). Specifically, they trim long videos into short clips given the ground-truth event boundaries E, train Vid2Seq on the short clips for sentence-level captioning, and concatenate the predictions on each clip to form paragraph-level captions. The proposal by Yang et al. (2023a) (named as "Vid2Seq-Concat" by us) has two weaknesses: (1) Vid2Seq-Concat simply divides the VPC task into video captioning on short clips and fails to model the inter-event dependence in each long video; (2) the video and ASR input of Vid2Seq-Concat is determined by the given event boundaries, which makes the system vulnerable when the event boundaries are noisy or absent. In comparison, our MVPC and MR-VPC schemes model all input modalities in an end-to-end manner, bringing two key advantages: (1) effective modeling of inter-event dependence in long videos; (2) no information loss in A and V when E is noisy. The experimental results on YouCook2 in Table 13 empirically validate the advantage of our proposals over Vid2Seq-Concat (Yang et al., 2023a)[5].

F Noise Besides Missing Modality

In § 5.2, we discuss five types of noise in auxiliary modalities. Here are the details of them:

- **Low-quality ASR**: In real-world scenarios, ASR systems may have hardware limitations, resulting in inferior ASR data compared to the ASR texts used during training generated by state-of-the-art ASR models. To simulate this situation, we replace the Whisper small.en model (244M parameters) with the tiny.en model (39M) and reduce the inference beam size from 5 to 1.
- **ASR Sentence Deletion**: To simulate the corruption of ASR data, we randomly delete 50% of all sentences in each test instance.
- **Event Deletion**: In order to simulate the corruption of event boundary data, we randomly delete 50% of the events in the event boundary data of each test instance.
- **Boundary Perturbation**: To introduce perturbations to the event boundaries, we add random uniform noise ranging from -5 to +5 units (percentage points) to each timestamp in the event boundaries of each instance.
- **Generated Boundary**: Considering that event boundaries predicted by models are more realistic noisy inputs than perturbed ground-truth boundaries, we leverage the PDVC (Wang et al., 2021) dense captioning model to generate event boundaries.

G Software and Hardware Requirements

We implement our code based on the PyTorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020) Python libraries. All experiments in this paper are conducted on a server with 8 NVIDIA A40 GPUs (48 GB memory per GPU).

H Qualitive Example

We present a qualitative case study in Figure 5 to highlight the strengths of our MVPC and MR-VPC models in the modality-complete setting. As shown, VidSeq and VLTinT baselines tend to produce hallucinations and predict concepts inconsistent with the video content. For example, although there is only one man moving and performing martial arts in the video, Vid2Seq predicts "*The men continue moving around one another*" and VLTinT generates "*another man is seen walking around him*". In contrast, our MVPC and MR-VPC models show almost no hallucinations. The generated captions are more accurate and closely aligned with the content of the video.

I Details of Human Evaluation

Three voluntary annotators, who are graduate students fluent in English, are asked to choose a caption that they deem more coherent with the

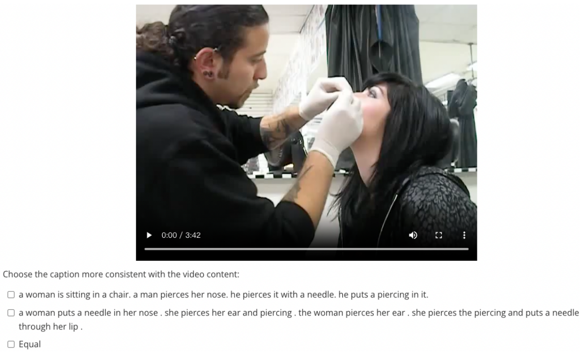
[5] We uniformly sample 15 frames for each clip in our implementation of Vid2Seq-Concat to keep the total input frames of each video close to 100.

Reference: A man is seen speaking to the camera and pans out into more men standing behind him. The first man then begins performing martial arts moves while speaking to he camera. He continues moving around and looking to the camera.
MVPC (Ours): A man is talking to the camera in a gym. Several martial arts are shown as he demonstrates them. A man is then seen performing several martial arts moves while the camera captures him from several angles.
MR-VPC (Ours): A man in a white t-shirt is talking to the camera. He is doing several martial arts moves on the mat. He does several kicks on the mat.
Vid2Seq: A man is seen speaking to the camera and leads into several shots of people performing martial arts moves. The men continue moving around one another while the camera captures their movements.
VLTinT: A man is seen speaking to the camera while standing in front of a large crowd. He is talking to the camera while another man is seen walking around him. He then does several martial arts moves while the camera captures his movements.

Figure 5: The captions produced by our models and baselines in the modality-complete setting on an ActivityNet Captions test sample (id: "bXdq2zI1Ms0"). The wrongly predicted concepts are highlighted in red by the author.

Model Predictions	**Modality-Complete Setting**	**Video-Only Setting**
Reference	pick the ends off the verdalago. combine lemon juice sumac garlic salt and oil in a bowl. chop lettuce and place it in a bowl. ···	
Vid2Seq	wash the leaves of verdolago. add lemon juice sumac crushed garlic salt and olive oil to a bowl and mix. ···	um, i'ma add some sea salt to the bowl. add some black pepper and mix it well. ···
MVPC	wash the pita bread slices. mix lemon juice sumac garlic salt and olive oil in a bowl. ···	tv.sv.svs.svv.svv on svvvm.svvm on svhvm on the svvm.
MR-VPC	wash the romaine lettuce leaves. add lemon juice sumac crushed garlic salt and olive oil to a bowl. ···	wash the romaine lettuce leaves. add lemon juice sumac crushed garlic salt and olive oil to a bowl. ···

Table 14: The predictions given by the models on a YouCook2 instance (id: "xHr8X2Wpmno") in the modality-complete setting (the second column) and the video-only setting (the third column). We only show the first two sentences of the predictions due to the limit of space and the degenerated predictions are highlighted in red.



Choose the caption more consistent with the video content:

- □ a woman is sitting in a chair. a man pierces her nose. he pierces it with a needle. he puts a piercing in it.
- □ a woman puts a needle in her nose . she pierces her ear and piercing . the woman pierces her ear . she pierces the piercing and puts a needle through her lip .
- □ Equal

Figure 6: The human annotation interface.

video content from a pair of model predictions or choose the "equal" option if they consider the two predictions to be equally good in terms of coherence. The data collection protocol is approved by an internal ethics review. We depict the layout of the annotation webpage in Figure 6.