A MULTI-MODAL BLIP-2 APPROACH FOR VIDEO CAPTIONING

CROSS-ATTENTION MECHANISM FOR MULTIMODAL FEATURE EXTRACTION

Auteurs

Antoine Brimont

Titus Zaharia, PhD thesis director

Partenaires

References

[1] Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, Li et al. in ICML, 2023.

[2] Pretrained image-text models are secretly video captioners, Zhang et al., 2025.

[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al., in NAACL 2019.

[4] Msr-vtt: A large video description dataset for bridging video and language, Xu et al., in CVPR 2016.

[5] Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. Zang et al., in ICCV, 2019.

[6] AudioCaps: Generating captions for audios in the wild, Kim et al., in NAACL-HLT, 2019.

INTRODUCTION

CONTEXT & OBJECTIVES

Video Captioning (VC) : derive a semantically pertinent textual description of a video segment.

THE AUDIO POTENTIAL

Audio remains underexplore while it as a rich source of information not only for speech but also for tonality, background noise or music.

THE CHALLENGE

Audio and visual modalities are closely related, but they have very different structures, which makes cross-modal reasoning challenging.

CURRENT LIMITATIONS

Existing works rely on either **separate paths** or **indiscriminate fusion**. Very few approaches attempt a seamless integration of both modalities.

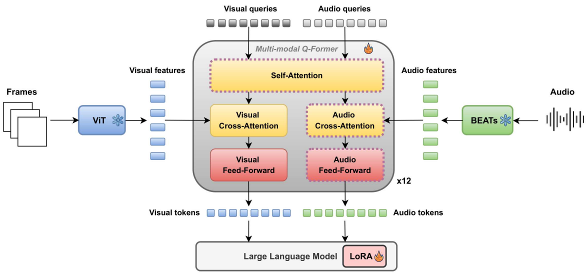

Figure 1 – Overview of the multi-modal BLIP-2 framework. Purple blocks denote our contribution.

METHOD

Q-FORMER

Introduced in BLIP-2, the Q-Former is architecturally derived from BERT. It consists of a stack of 12 transformer layers in which learnable queries extract relevant information from visual features through cross-attention modules.

MULTIMODAL BLIP-2

Building on **BLIP-2**[1] and **PIT-VC**[2] frameworks, our approach (**Figure 1**) samples 16 video frames and extract features via **ViT** and **BEATs**. Visual features are concatenated and then fed alongside audio features to a **multi-modal Q-Former**, producing a compact set of audio-visual tokens that guide a LLM, T5-Flan-XL, for caption generation. Furthermore, we leverage **effective transfer learning** from BLIP-2 to enhance training efficiency and performance.

MULTIMODAL Q-FORMER

Our multimodal Q-Former (**Figure 2**) employs **separate cross-attention** streams for audio/visual features but enables interaction through **shared self-attention module**. This compresses large feature sets into **a compact set of 48 tokens** (32 visual, 16 audio) while preserving modality-specific details and enabling rich cross-modal interactions.

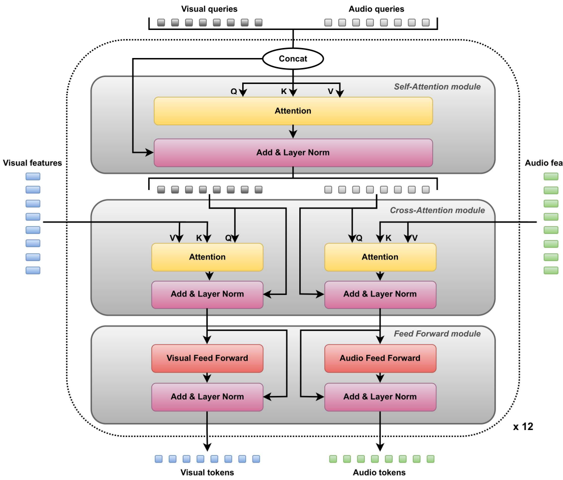

Figure 2 – Details of the multi-modal Q-Former.

EXPERIMENTAL RESULTS

Evaluation on three publically available benchmark datasets: **MSRVTT**[4] and **Latest-VATEX**[5], and **AudioCaps**[6],

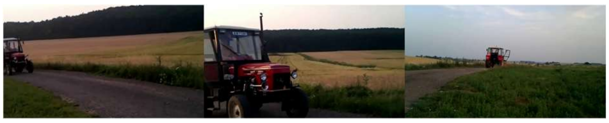

Figure 3 – Qualitative result on AudioCaps dataset.

GT: A tractor driving by as a car horn honks while wind blows into a microphone.

PIT-VC (Baseline): A vehicle engine idling and humming.

Ours: A vehicle horn honks followed by wind blowing into a microphone.

	MSR-VTT					Latest-VATEX					Model Info		
Model	C	M	R	B	S	C	M	R	B	S	Audio	#PT	Size
mPLUG-2	71.4	32.9	67.2	51.4	8.9	–	–	–	–	–	No	2.5M	1.1B
VALOR	74.0	32.9	68.0	54.4	–	82.3	28.1	56.7	45.5	13.9	Yes	1.18M	800M
VAST*	78.0	–	–	**56.7**	–	86.6	28.7	**57.3**	**46.9**	14.6	Yes	27M	1.4B
PIT-VC	79.5	34.2	68.3	52.4	–	84.2	28.2	56.4	44.1	14.5	No	0	4.1B
Separate	77.1	33.9	68.5	53.6	8.8	85.1	28.5	56.6	44.6	14.6	Yes	0	4.2B
Ours	**80.1**	**35.0**	**69.5**	55.0	**9.3**	**86.8**	**28.8**	56.7	44.2	**15**	Yes	0	4.2B

Table 1 – Quantitative comparison on the MSR-VTT and VATEX datasets. The asterisk (*) denotes models utilizing speech as an additional input modality.

	AudioCaps					Model Info		
Model	C	M	R	B	S	Visual	#PT	Size
VAST*	78.1	24.7	–	–	–	Yes	0	1.4B
AutoCap*	83.2	25.3	–	–	18.2	No	400K	1.25B
LOAE	81.6	26.7	–	–	19.3	No	400K	7B
SLAM-ACC	**84.1**	**26.8**	–	–	**19.4**	No	400K	7B
PIT-VC	61.9	20.8	44.3	20.7	14.2	Yes	0	4.1B
Separate	76.0	25.1	49.6	27.1	18.3	Yes	0	4.2B
Ours	82.5	26.0	**51.1**	**27.7**	18.6	Yes	0	4.2B

Table 2 – Quantitative comparison on the AudioCaps test set. The asterisk (*) denotes models utilizing speech as an additional input modality.

Without pre-training, our model achieves SOTA performance on visually centered datasets and demonstrates strong results on AudioCaps.

Journée des doctorants 2026

February 2026

Contact antoine.brimont@telecom-sudparis.eu