Companion Paper on Multi-modal BLIP-2: Q-Former Variants and Metric Robustness

Antoine Brimont[1 [0009−0002−8702−7961]], Titus Zaharia[1 [0000−0002−6589−1241]], and Ruxandra Tapu[1 [0000−0003−3170−4150]]

SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France

Abstract. This paper serves as a companion study to our ICPR 2026 paper[1], "A Multi-modal BLIP-2 Approach for Video Captioning". Its goal is to provide a deeper analysis of our proposed multi-modal adapter for Video-Language Models (VLMs), the multi-modal Q-Former, and of the architectural factors that govern its performance. We first present a comprehensive comparison of several variants of the multi-modal Q-Former, isolating the effect of the number of cross-attention layers, the number of audio queries, the organization of the audio and visual cross-attention modules (alternating versus aligned configurations), and the initialization of the model from the original BLIP-2 Q-Former weights. These variants are evaluated on the MSR-VTT and AudioCaps benchmarks under a lightweight training protocol, and are compared against unimodal and separate-Q-Former baselines. In the second part, we focus on the stability of comparisons, with particular emphasis on the impact of hyperparameters. While the original paper reports results only under the best-performing generation settings, we conduct an extensive study of 25 decoding configurations obtained by jointly varying the length and repetition penalties. This analysis clarifies the influence of decoding hyperparameters on caption quality and quantifies whether the observed improvements are primarily due to the proposed architecture or to the decoding strategy itself. Finally, we examine how evaluation metrics respond to these hyperparameters, showing that n-gram-based metrics (BLEU-4, ROUGE-L, CIDEr) and the scene-graph-based SPICE exhibit opposite sensitivities. As a result, no single decoding configuration jointly optimizes all metrics.

Keywords: Video Captioning · Multi-Modal Learning · Audio-Visual Understanding

1 Introduction

In our previous work, A Multi-modal BLIP-2 Approach for Video Captioning [2], we proposed a novel multi-modal extension of BLIP-2 [9], inspired by the PIT-VC framework [17], which extends the Q-Former architecture to video understanding

[1] Original code available at `https://github.com/ABrimont/Multi-modal-BLIP-2`

while leveraging BLIP-2's pretrained weights. In our design, frames are uniformly sampled from the video and encoded by a frozen ViT [5], while the accompanying audio is encoded by BEATs [3]. Visual and audio features are then compressed by the multi-modal Q-Former into a compact set of tokens passed to the Flan-T5-XL [12] text decoder for caption generation. Its key novelty lies in introducing early audio-visual interactions within the Q-Former, through modality-specific query tokens and dedicated cross-modal reasoning modules, rather than deferring fusion to the language model. This design enables the model to capture complementary information from both modalities while preserving their individual semantic representations. Despite being trained with a lightweight strategy and without any video-text pre-training, the proposed method achieves competitive, and in several settings state-of-the-art results on the MSR-VTT [16], VATEX [14], and AudioCaps [7] benchmarks. While these results validate the effectiveness of the proposed architecture, a more comprehensive analysis of the multi-modal Q-Former design remains necessary. In particular, the contributions of its individual architectural components and cross-modal interaction mechanisms have not yet been thoroughly investigated.

Furthermore, we observe that decoding hyperparameter selection plays a critical role in achieving optimal performance. Since benchmark scores are often separated by only a few percentage points, suboptimal decoding configurations can substantially affect the final results and may prevent a model from reaching state-of-the-art performance. To better understand this phenomenon, we conduct our analysis in two complementary stages. First, we compare all architectural variants using their respective best decoding hyperparameter configurations, allowing us to isolate the effect of the proposed architectural modifications. Second, we evaluate every model under 25 different decoding configurations obtained by jointly varying the length and repetition penalties. For each architecture, we report both the mean performance and the standard deviation across all configurations. This analysis enables us to assess the robustness of each model to decoding hyperparameter variations and to determine whether the observed performance gains originate primarily from the proposed architecture or from favorable decoding settings. Finally, we reverse the perspective by analyzing the sensitivity of each evaluation metric to the decoding hyperparameters, investigating whether all metrics share the same optimal configuration or whether different metrics favor different decoding strategies.

We conduct our study on the MSR-VTT and AudioCaps benchmarks, while leaving the VATEX dataset for future work in order to limit computational requirements. To enable a fair comparison across a large number of architectural variants and decoding configurations, all trainings are performed under a lightweight protocol, which involves a single training stage.

2 Impact of the Multi-modal Q-Former Structure

In this section, we investigate the impact of the main design choices underlying the proposed multi-modal Q-Former. We perform an extensive ablation study to

better understand the contribution of each architectural choice and identify the factors that most strongly influence performance.

2.1 Architectural Variants of the Multi-modal Q-Former

We evaluate several variants of the proposed architecture, each designed to isolate a specific design choice. Unless otherwise stated, all other architectural components are kept identical, and all variants are trained under comparable settings. Except for the *Random Initialization variant*, all models are initialized as described in our original paper. Specifically, the visual and audio inputs are processed by the frozen pretrained ViT and BEATs encoders, and the text decoder retains its pretrained weights. Within the multimodal Q-Former, the self-attention modules, the visual cross-attention layers, and the visual feed-forward layers are initialized from the pretrained BLIP-2 Q-Former weights, while all remaining components are initialized randomly.

Aligned Cross-Attention (6 Layers). We first consider the architecture proposed in our original paper, denoted as *Aligned-6*. Following the publicly released BLIP-2 weights, audio and visual cross-attention modules are inserted in even- numbered layers. Consequently, each modality performs six cross-attention operations across the twelve layers of the Q-Former. At each selected layer, visual queries attend to visual features while audio queries attend to audio features. The model uses 32 visual queries and 16 audio queries, reflecting the assumption that audio information is less dense than visual information. This configuration was initially chosen to leverage the pre-trained BLIP-2 weights while maintaining a lightweight design.

Aligned Cross-Attention (12 Layers). To investigate whether more frequent interactions with the frozen encoders are beneficial, we consider an *Aligned-12* variant in which both audio and visual cross-attention modules are inserted at every transformer layer. Consequently, each modality performs twelve cross-attention operations instead of six, allowing more frequent feature extraction and modality-specific refinement throughout the network.

Alternating Cross-Attention. While the previous configurations apply audio and visual cross-attention within the same transformer layers, we also investigate an *Alternating* strategy. In this variant, cross-attention modules are distributed across layers in an alternating fashion: visual cross-attention is applied in even-numbered layers, while audio cross-attention is applied in odd-numbered layers. As a result, six layers are dedicated to visual cross-attention and six layers to audio cross-attention, yielding twelve cross-attention operations in total. This design encourages a progressive exchange of information between modalities throughout the depth of the network.

Increased Audio Query. Our original architecture employs 32 visual queries and 16 audio queries, as we suppose a higher information density in visual representations. To evaluate the influence of query capacity, we introduce a variant using 32 queries for both modalities. This configuration increases the representational capacity of the audio branch and enables a more balanced allocation of learnable queries between audio and vision.

Separate Q-Formers. As a first baseline, we consider the *Separate* architecture introduced in our original paper. This variant follows the design adopted by Video-LLaMA [4] and employs two independent Q-Formers, one dedicated to visual features and one to audio features. Each Q-Former processes its corresponding modality independently, with no interaction between audio and visual representations during feature extraction. Cross-modal alignment is therefore deferred to the language model, which receives the visual and audio tokens only after the compression stage. This baseline allows us to evaluate the benefits of introducing early audio-visual interactions within the Q-Former itself.

Unimodal Q-Former. As a second baseline, we consider the original *Unimodal* architecture introduced in PIT-VC. This model processes only the visual modality and relies on a single Q-Former initialized from the pre-trained BLIP-2 weights. The Q-Former compresses the visual features extracted from multiple video frames into a fixed set of visual tokens, which are subsequently passed to the language model for caption generation. Unlike the proposed multi-modal architecture, no audio information is incorporated and no cross-modal interactions are performed. This baseline allows us to quantify the contribution of audio integration independently of the proposed multi-modal Q-Former design.

Random Initialization. Finally, we assess the importance of transfer learning from BLIP-2 through a *Random Initialization* variant. Unlike the original model, which initializes the self-attention, visual cross-attention, and visual feed-forward layers using pre-trained BLIP-2 weights, all Q-Former parameters are initialized randomly, while the text decoder remains pre-trained. This experiment quantifies the contribution of BLIP-2 pre-training and helps determine whether the observed performance gains primarily originate from the proposed architectural design or from transferred image-text knowledge.

2.2 Training settings

Following our original work, and in order to limit computational costs, all architectural variants are evaluated on the MSR-VTT and AudioCaps datasets. All experiments follow the first training stage described in our original paper, with the sole modification that the Flan-T5-XL decoder is fine-tuned using LoRA [6] with a rank of 8. The multi-modal Q-Former and the LoRA adapters are jointly optimized using the standard cross-entropy objective, while all remaining training hyperparameters are kept identical to those of the original stage-1 training.

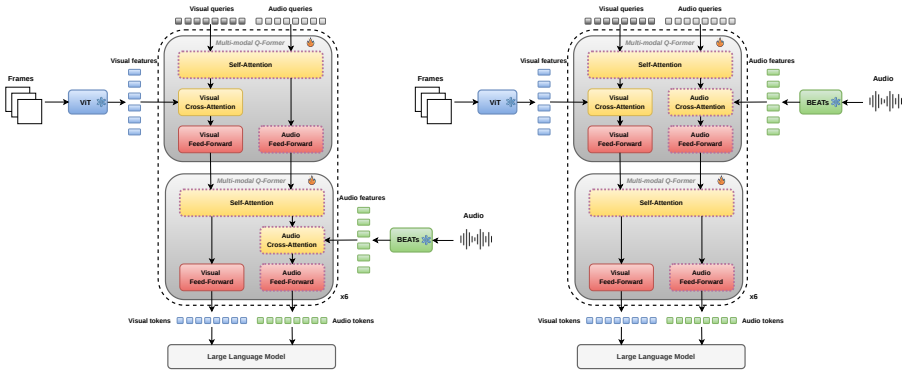

Fig. 1. Details of the Alternating Cross-Attention (left) and the Aligned-6 (right) variants. Purple blocks denote our contributions.

This protocol enables us to isolate the impact of the proposed architectural modifications while maintaining a reasonable computational budget.

Since the objective of this study is to compare architectural variants rather than maximize benchmark performance, we omit the second training stage based on self-critical sequence training (SCST).

2.3 Metrics

We evaluate the quality of the generated captions using five standard captioning metrics. BLEU-4 [11] measures the precision of 4-gram overlaps between generated and reference captions. ROUGE-L [10] evaluates the longest common subsequence, capturing sentence-level similarity while preserving word order. METEOR [8] extends lexical matching by considering stemming and synonymy, providing a more semantically informed comparison. CIDEr [13] computes a TF-IDF-weighted consensus score over n-grams, emphasizing informative words that are consistently present across multiple reference captions. Finally, we report SPICE [1], which evaluates semantic consistency by comparing scene graphs extracted from the generated and reference captions. Together, these metrics provide complementary perspectives on lexical overlap, semantic similarity, and content relevance.

While the retained metrics are widely used to evaluate video captioning models, they remain strongly influenced by the length of the generated outputs and specific word choices, rather than by the underlying semantic meaning of the sentences. This limitation is further discussed in Section 3.

2.4 Experimental comparison of the architectural variants

Table 1 reports the best score obtained by each variant, selected according to the CIDEr metric. The optimal decoding hyperparameters are determined through a

Table 1. Best performances of the Q-Former variants on MSR-VTT and AudioCaps. Bold indicates the best performance. Abbreviations: C (CIDEr), M (METEOR), R (ROUGE-L), B (BLEU-4), S (SPICE).

	MSR-VTT					**AudioCaps**				
Variant	C	M	R	B	S	C	M	R	B	S
Unimodal	73.6	33	67.2	51.9	8.7	62.1	20.4	44	21.4	15
Separate	73.7	33.2	67.2	51.3	8.8	73.6	24.1	48.8	25.1	18
Aligned-6 (Original)	76.9	34.1	**68.4**	53	**9.2**	80.6	26.4	50.5	27.4	**19.9**
Aligned-12	75.1	33.9	68.1	53.6	**9.2**	**82.1**	25.6	50.9	27.6	19.3
Alternating	**77.1**	**34.2**	**68.4**	**55.3**	8.9	77.9	25.4	50.3	28.2	18.9
Increased Audio Query	74.9	34.1	68.4	53.1	**9.2**	**82.1**	**26.7**	**51.3**	**28.5**	**19.9**
Random Initialization	65.2	32	66.5	50.2	7.8	76.7	24.8	49.6	27.2	18.1

grid search over repetition and length penalties, with values in 0.7, 0.85, 1.0, 1.15, 1.3. The repetition penalty rescales the logits of previously generated tokens, reducing the probability of repetition, encouraging more diverse outputs when increased. The length penalty adjusts the beam search score by normalizing sequence log-probabilities with respect to length, thereby controlling the trade-off between shorter and longer captions. For decoding, we use beam search with a beam size of 5, while constraining the generated captions to a minimum of 5 and a maximum of 50 tokens.

For both benchmarks, all multi-modal variants initialized from pretrained BLIP-2 weights consistently outperform both the unimodal and separate variants. These results reinforce the conclusions of our previous work, confirming that introducing early cross-modal interactions within the Q-Former provides a significant advantage for video captioning. Conversely, randomly initializing the Q-Former leads to rapid overfitting and poor performance, further supporting the findings of PIT-VC [17] and emphasizing the crucial role of transfer learning from pretrained vision–language models. Overall, all four architectures combining early cross-modal interactions with pretrained initialization achieve competitive performance. The remaining gap with the results reported in our original paper can largely be attributed to our lightweight training protocol. In particular, for MSR-VTT, we omit the SCST fine-tuning stage, which is essential for fully exploiting the dataset's dense annotations (20 captions per video).

Some differences can be observed between the two datasets. On MSR-VTT, where visual information is substantially more informative than audio cues, the lightest variants, *Aligned-6* and *Alternating*, achieve the best performance, likely because they better preserve and exploit the visual knowledge transferred from the pretrained BLIP-2 weights.

In contrast, on AudioCaps, the larger variants, *Aligned-12* and *Increased Audio Query*, achieve better results. In particular, increasing the number of audio queries to 32 leads to a substantial performance improvement, with gains up to 2% for CIDEr over the original architecture. This variant consistently outperforms all other counterparts on AudioCaps and achieves a new state-of-the-art SPICE score, despite relying on a remarkably simple training protocol.

Table 2. Ablation study of the Q-Former variants on MSR-VTT and AudioCaps. Each cell reports the mean with ± standard deviation; bold indicates the best mean per column. Abbreviations: C (CIDEr), M (METEOR), R (ROUGE-L), B (BLEU-4), S (SPICE).

	MSR-VTT					AudioCaps				
Variant	C	M	R	B	S	C	M	R	B	S
Unimodal	70.9 ± 2.7	32.7 ± 0.5	66.2 ± 1.1	49 ± 3.5	8.9 ± 0.4	60.4 ± 2	21.1 ± 0.5	43.6 ± 0.5	19.7 ± 1.7	15.2 ± 0.2
Separate	71.2 ± 2.3	32.9 ± 0.4	66.6 ± 0.9	49.6 ± 3	8.8 ± 0.4	70.6 ± 3.1	24.5 ± 0.6	48.3 ± 0.9	23.9 ± 2.1	18 ± 0.5
Aligned-6 (Original)	71.9 ± 3.7	33.3 ± 0.7	66.4 ± 1.6	46.6 ± 4.7	**9.7** ± 0.4	78.4 ± 1.7	26 ± 0.4	50.2 ± 0.4	27.2 ± 1.9	19.5 ± 0.4
Aligned-12	71.2 ± 2.7	33.3 ± 0.7	67 ± 1.2	49.6 ± 4	9.4 ± 0.3	**80.5** ± 1.3	25.5 ± 0.5	50.6 ± 0.3	26.5 ± 1.4	19.2 ± 0.5
Alternating	**74.1** ± 2.7	33.7 ± 0.5	**67.8** ± 1.2	50.1 ± 4.1	9.4 ± 0.3	76.4 ± 0.4	25.3 ± 0.4	50 ± 0.3	27 ± 1.7	19 ± 0.5
Increased Audio Query	72.8 ± 2.1	**33.8** ± 0.3	**67.8** ± 1	**51.4** ± 3.3	9.2 ± 0.4	**80.5** ± 1.5	**26.2** ± 0.4	**51.1** ± 0.4	**27.7** ± 1.3	**19.6** ± 0.5
Random Initialization	62.5 ± 2.4	31.7 ± 0.5	65.1 ± 1.2	45.8 ± 3.7	8.2 ± 0.3	73.9 ± 2.5	24.4 ± 0.4	49.1 ± 0.7	24.9 ± 2.1	18 ± 0.5

The obtained results further confirm our findings on the importance of early cross-modal reasoning, even under a lightweight training protocol. Although none of the proposed variants consistently outperforms the others, our results suggest that neither introducing cross-attention at every layer nor adopting an alternating interaction scheme provides a clear advantage over simpler designs. Instead, allocating an appropriate number of queries to each modality according to its information density appears to be more important. Overall, these observations indicate that the precise placement and density of cross-attention connections are less critical than combining early cross-modal interactions with a pretrained Q-Former initialization.

Further work could investigate the benefits of larger query sets for both visual and audio modalities, particularly in scenarios requiring denser feature representations, such as long-form videos, fine-grained video understanding, or precise visual question answering. Such settings may benefit from an increased number of queries, allowing the Q-Former to capture a richer and more diverse set of multimodal features. In addition, exploring more recent large language models could further improve reasoning and generation capabilities. In this work, we deliberately retain FLAN-T5-XL to preserve the pretrained BLIP-2 initialization and ensure a fair evaluation of the proposed Q-Former architectures, rather than introducing improvements stemming from a stronger language model.

3 Sensitivity to the Choice of Hyperparameters

Following the setup already introduced in the previous section, Table 1 reports the average scores and standard deviations of each variant on both datasets over the 25 decoding configurations obtained through the grid search over repetition and length penalties, in order to assess the robustness of each model to decoding hyperparameter variations rather than its performance under a single optimal setting. In this section, we analyze the sensitivity of both the model variants and the evaluation metrics to these hyperparameters, and we determine whether the conclusions drawn from the best-performing configurations still hold on average across the whole grid.

3.1 Robustness to multiple hyperparameter settings

To provide a more comprehensive analysis of the reported results, Figure 2 and Table 2 illustrate the impact of decoding hyperparameters on the performance of each model variant.

The overall ranking of the models remains consistent with the best results reported in Table 1 on both datasets. In particular, the four multimodal Q-Former variants consistently outperform both the unimodal and separate-modality baselines on average across the different decoding hyperparameter configurations. This indicates that the conclusions drawn from the best-performing settings are robust and do not depend on a particular choice of decoding parameters. Although hyperparameter tuning can influence the performance of closely competing architectures, it does not alter the overall ranking or the various architectural variants retained in our study.

Nevertheless, the box plots clearly show that inappropriate decoding hyperparameters can substantially degrade performance, leading to scores that are significantly lower than those obtained with the optimal configuration. This highlights the importance of carefully tuning decoding parameters, not only to achieve the best possible performance but also to ensure fair comparisons between competing architectures and model variants. Furthermore, the sensitivity to these hyperparameters can lead to suboptimal model selection, particularly when validation is performed on relatively small evaluation sets, where performance estimates are inherently more affected by the variance induced by decoding settings.

Another interesting observation is that the variance induced by decoding hyperparameters is consistently larger on MSR-VTT than on AudioCaps, despite the much larger MSR-VTT evaluation set (2990 vs. 975 samples), which would typically be expected to yield more stable performance estimates. Further investigation is needed to explain this behavior. One possible hypothesis is that the dense annotation setting of MSR-VTT (approximately 20 captions per training video, compared with a single caption per sample in AudioCaps) encourages the model to learn a broader distribution of valid descriptions, making the gener-

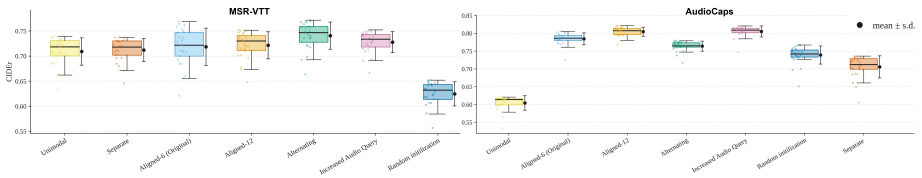

Fig. 2. Box plots of the performance achieved by each variant across all grid search hyperparameter configurations for MSR-VTT (left) and AudioCaps (right).

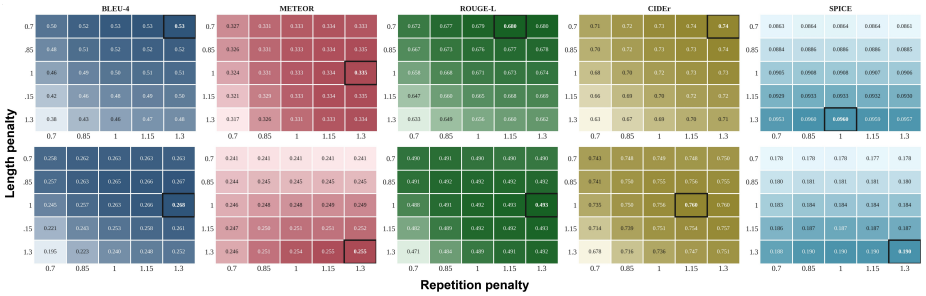

Fig. 3. Heatmaps showing the relative performance of each evaluation metric across the grid search over length and repetition penalties on MSR-VTT (top) and AudioCaps (bottom). Scores are normalized by the mean performance over all decoding configurations. Darker shades indicate better performance, while black boxes highlight the best-performing configuration for each metric.

ated caption, and consequently the evaluation metric, more sensitive to decoding hyperparameters.

3.2 Metrics sensitivity

From the perspective of the evaluation metrics, we further investigate whether they share a common optimal decoding configuration or instead favor different hyperparameter settings. Figure 3 presents heat maps of the performance obtained by each metric across the different combinations of length and repetition penalties. While this analysis focuses on these two hyperparameters for visualization purposes, a more comprehensive study could also investigate the impact of other decoding parameters, including beam size and output length constraints, but also of other decoding strategies. This was notably made on Large Language Model [15].

First, we observe that the heat maps differ across datasets. Unsurprisingly, these decoding hyperparameters are data-dependent, varying with both the data used for training and that used for evaluation. On MSR-VTT, a general trend emerges in favor of shorter generated sequences combined with higher repetition penalties, whereas on AudioCaps, better results are obtained with moderate length values and higher repetition penalties. In both cases, results show that higher repetition penalties should be privileged.

Despite these differences between datasets, consistent trends emerge across evaluation metrics. BLEU-4, ROUGE-L, and CIDEr exhibit smooth and highly similar response surfaces, indicating that they favor comparable decoding strategies. Among them, BLEU-4 is by far the most sensitive to decoding hyperparameters, with performance varying by up to 30% across the explored grid. In contrast, METEOR is considerably more stable, with variations limited to approximately 5% over the same configurations. SPICE exhibits a distinct behavior, achieving its highest scores for large length penalty values and gradually

decreasing as the length penalty is reduced. Consequently, no single decoding configuration simultaneously optimizes all evaluation metrics, highlighting the inherent trade-offs involved in selecting decoding hyperparameters.

These differences are consistent with the underlying design of the metrics. BLEU-4, ROUGE-L, and CIDEr are all primarily driven by lexical overlap with reference captions and therefore respond similarly to changes in decoding hyperparameters that mainly affect surface form, such as length and repetition penalties, which may explain their close agreement. Furthermore, BLEU-4 relies on exact 4-gram matching and is thus highly sensitive to small lexical or syntactic variations, even when the generated caption preserves the intended meaning. METEOR incorporates stemming and synonym matching, making it more robust to surface-level variations and consequently less sensitive to decoding choices. In contrast, SPICE evaluates captions through semantic scene graphs rather than lexical overlap. As a result, it is comparatively insensitive to syntactic variations such as length changes and instead primarily penalizes semantic omissions and hallucinations. These complementary behaviors explain why different metrics favor distinct decoding configurations.

4 Conclusions and perspectives

This companion paper presents a comparative analysis of multi-modal Q-Former variants for video captioning, along with a study of decoding hyperparameters and evaluation metrics.

Our study confirms that the two decisive factors behind the observed gains are early cross-modal interactions within the Q-Former and transfer learning from BLIP-2: all variants combining both consistently outperform the unimodal and separate baselines, while random initialization collapses. In contrast, the precise placement and density of cross-attention connections matter less than allocating queries to each modality according to its information density, as shown by the *Increased Audio Query* variant reaching a new state-of-the-art SPICE score on AudioCaps.

Across the 25 decoding configurations considered for evaluation, the overall ranking remains stable and each model's best score correlates with its average, confirming that our conclusions do not depend on a particular decoding choice. However, suboptimal settings can severely degrade performance, stressing the need for careful tuning to ensure fair comparisons. Finally, the metrics exhibit opposite sensitivities: lexical-overlap metrics (BLEU-4, ROUGE-L, CIDEr) favour short captions while SPICE rewards longer ones, so that no single decoding configuration jointly optimizes all metrics.

Future work could pursue several complementary directions. A natural next step would be to train the multi-modal Q-Former on larger corpora and more challenging tasks, such as longer videos and richer audio-visual interactions, although dense audio-visual datasets remain scarce and difficult to obtain. Replacing the Flan-T5-XL decoder with a more recent language model could further improve caption quality, provided a fair evaluation of the underlying architecture

is preserved. Finally, extending the hyperparameter analysis to a broader set of datasets and to additional decoding parameters, such as beam size, would yield stronger and more general conclusions.

Acknowledgment

This research work has been carried out within the framework of the AITV joint laboratory between Télécom SudParis and France Télévisions.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European conference on computer vision. pp. 382–398. Springer (2016)
2. Brimont, A., Zaharia, T., Tapu, R.: A multi-modal blip-2 approach for video captioning. In: Proceedings of the 28th International Conference on Pattern Recognition (ICPR). Lecture Notes in Computer Science, Springer (2026)
3. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., et al.: Beats: Audio pre-training with acoustic tokenizers. arXiv preprint arXiv:2212.09058 (2022)
4. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
5. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
7. Kim, C.D., Kim, B., Lee, H., Kim, G.: AudioCaps: Generating captions for audios in the wild. In: Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. pp. 119–132 (2019), https://aclanthology.org/N19-1011/
8. Lavie, A., Agarwal, A.: Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation. p. 228–231 (2007)
9. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
10. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002), https://aclanthology.org/P02-1040/
12. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
13. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)

14. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proc. of the IEEE/CVF international conference on computer vision. pp. 4581–4591 (2019)
15. Wiher, G., Meister, C., Cotterell, R.: On decoding strategies for neural text generators. Transactions of the Association for Computational Linguistics **10**, 997–1012 (2022). https://doi.org/10.1162/tacl_a_00502, https://aclanthology.org/2022.tacl-1.58/
16. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
17. Zhang, C., Jian, Y., Ouyang, Z., Vosoughi, S.: Pretrained image-text models are secretly video captioners. arXiv preprint arXiv:2502.13363 (2025)