Towards Visually Grounded Multimodal Summarization via Cross-Modal Transformer and Gated Attention

Abid Ali* **and Diego Mollá-Aliod and Usman Naseem**
School of Computing, Macquarie University, Sydney, Australia
abidmeeraj@gmail.com, diego.molla-aliod@mq.edu.au, usman.naseem@mq.edu.au

Abstract

Multimodal summarization requires models to jointly understand textual and visual inputs to generate concise, semantically coherent summaries. Existing methods often inject shallow visual features into deep language models, leading to representational mismatches and weak cross-modal grounding. We propose a unified framework that jointly performs text summarization and representative image selection. Our system, SPeCTrA-Sum (Sampler Perceiver with Cross-modal Transformer and gated Attention for Summarization), introduces two key innovations. First, a Deep Visual Processor (DVP) aligns the visual encoder with the language model at corresponding depths, enabling hierarchical, layer-wise fusion that preserves semantic consistency. Second, a lightweight Visual Relevance Predictor (VRP) selects salient and diverse images by distilling soft labels from a Determinantal Point Process (DPP) teacher. SPeCTrA-Sum is trained using a multi-objective loss that combines autoregressive summarization, cross-modal alignment, and DPP-based distillation. Experiments show that our system produces more accurate, visually grounded summaries and selects more representative images, demonstrating the benefits of depth-aware fusion and principled image selection for multimodal summarization.

1 Introduction

In today's information-rich environment, users are often overwhelmed by the need to quickly process large volumes of multimodal content. This includes news articles with embedded images, blog posts featuring photo galleries, or technical reports combining diagrams and captions. While traditional summarization models focus exclusively on textual compression, real-world documents often span both visual and textual modalities (Zhu et al., 2018). As a result, multimodal summarization, the task of generating concise summaries conditioned on both text and images, has emerged as an increasingly important research direction. Figure 1 shows an example of multimodal summarization with multimodal output.

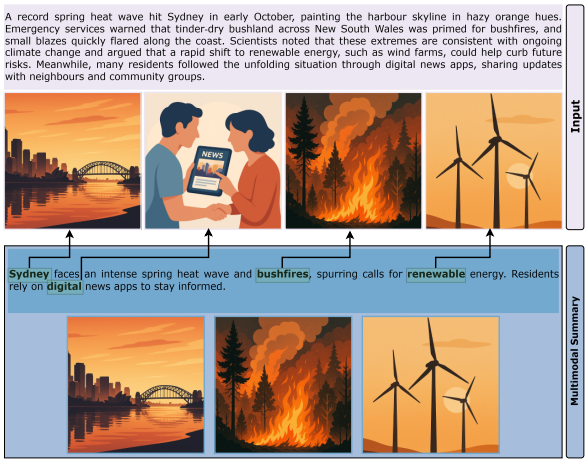


Figure 1: An Illustration of a Multimodal System with Multimodal Output.

Recent surveys emphasize that multimodal summarization must effectively integrate heterogeneous inputs from different modalities (Atharva et al., 2023; Zhang et al., 2023; Jangra et al., 2023). However, despite substantial progress in multimodal learning, many models continue to exhibit fundamental limitations in cross-modal integration, largely due to a persistent semantic gap between visual and textual representations (Zhu et al., 2018).

In practice, visual features are often only shallowly fused into deep language models, resulting in weak semantic alignment between modalities (Liu and Liu, 2025). Consequently, image representations frequently fail to capture deeper textual abstractions (Chen et al., 2025), limiting effective cross-modal grounding. For example, models may struggle to reliably associate visual content with its corresponding textual descriptions (Li et al., 2025b). As a result, vision and language remain only loosely coupled in many existing systems,

*Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2026, pages 29032–29047

underscoring the need for architectures that more effectively bridge this representational divide.

Another key challenge is selecting images that meaningfully contribute to the summary. Source documents often include many images, some of which are redundant or peripheral. Thus, identifying a representative, non-redundant subset is essential. This involves a dual goal: selecting images that are both individually relevant and collectively diverse. In related areas like image summarization and keyframe extraction, Determinantal Point Processs (DPPs) have proven effective in modeling this relevance–diversity trade-off (Celis et al., 2018; Cho et al., 2019).

Motivated by the inherent coupling between cross-modal alignment and visual subset selection, we propose a unified architecture that jointly optimizes abstractive summary generation and the selection of image subsets that best support the generated text. Our key contributions are:

- **Joint multimodal summarization and image selection:** We model image selection as an integral component of the summarization process rather than a post-hoc step. Specifically, we introduce the Visual Relevance Predictor (VRP), which jointly learns to generate a summary Y and select a representative image subset $I^* \subseteq I$ from the input text X and image set I, enabling tighter alignment between textual and visual outputs.
- **Knowledge-distilled image selector with DPP supervision:** To move beyond heuristic image scoring, we employ a DPP-based teacher that jointly models text–image relevance and inter-image diversity to produce soft inclusion probabilities. These pseudo-labels are used to train the VRP, enabling it to efficiently select relevant and non-redundant image subsets at inference time.
- **Multi-objective training for cross-modal alignment and diversity-aware selection:** We train the model with a unified objective that combines (i) an autoregressive summarization loss, (ii) a cross-modal alignment loss between generated text and visual features, and (iii) a distillation loss derived from the DPP teacher. Together, these objectives encourage the model to jointly optimize textual quality, visual grounding, and diversity-aware image selection.

We evaluate our method on the MSMO dataset (Zhu et al., 2018) and demonstrate that it (i) improves standard summarization metrics over baselines without image selection, (ii) produces more visually grounded summaries with higher image–text relevance, and (iii) selects image subsets that better balance relevance and diversity compared to heuristic or text-only approaches.

2 Related Work

2.1 Text and Multimodal Summarization

Early abstractive summarization models framed the summarization task as sequence modeling with attention. Rush et al. (2015) introduced a neural attention-based model for sentence-level summarization, outperforming extractive baselines. Follow-ups (Nallapati et al., 2016; See et al., 2017, for example) improved factuality and coverage using pointer-generator mechanisms and copy-aware decoding.

Multimodal summarization extends this task to include visual inputs such as images or videos. Zhu et al. (2018) formally defined the multimodal setting and proposed a joint model for text generation and image selection, along with the MSMO dataset. Later works explored multimodal attention to improve informativeness and grounding, showing that integrating image features via attention can enhance summary quality (Chen and Zhuge, 2018; Li et al., 2018).

2.2 Multimodal Fusion and Alignment in Vision-Language Models

Early Vision-Language (VL) models like ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) used dual-stream encoders with cross-attention to learn joint multimodal representations, improving performance on tasks like VQA. More recent models, such as Flamingo (Alayrac et al., 2022), integrate pretrained vision and language modules using gated cross-attention, enabling few-shot generation from interleaved image–text inputs.

Several studies (Cao et al., 2020; Bugliarello et al., 2021) show that deeper layers tend to prioritize textual representations while paying limited attention to visual inputs, motivating the need for more effective layer-aligned fusion strategies.

2.3 Image Selection and Visual Importance

Multimodal summarization requires identifying which images meaningfully support the generated summary. Including all available images can dilute relevance or introduce redundancy, motivating the need for selective conditioning. Liang et al.

(2023) improve visual encoder utility via auxiliary objectives like vision-to-summary prediction and masked image modeling. CFSum (Xiao et al., 2023) filters uninformative images through a preselection module and modulates attention with visual complement units. DIUSum (Xiao et al., 2024) further predicts image usefulness and dynamically gates visual reliance during decoding. Together, these approaches highlight the benefits of targeted visual integration for content quality and efficiency.

2.4 Knowledge Distillation and Multi-Objective Learning

Knowledge Distillation (KD) (Hinton et al., 2015) trains compact models to mimic larger ones via softened output distributions. In multimodal settings, KD has enabled cross-modal supervision; for instance, Gupta et al. (2016) transferred knowledge from labeled RGB to depth/optical flow streams using paired data.

For VL pretraining, ALBEF (Li et al., 2021) combined contrastive alignment with momentum-based self-distillation to learn from noisy web pairs. More recent work explores distilling large CLIP models into smaller ones (e.g., CLIP-KD (Yang et al., 2024; Pei et al., 2023)), extending KD to video–language tasks.

In summary, prior work has advanced multimodal summarization through improved fusion strategies, image filtering, and cross-modal alignment. However, several challenges remain. Image selection is often heuristic or treated as a separate stage from text generation, limiting joint optimization and alignment. Moreover, most approaches prioritize relevance while overlooking diversity within the selected image set, resulting in redundant or less informative visual summaries. To address these limitations, we propose a unified framework that integrates gated, depth-aware fusion, distillation from a diversity-aware teacher, and multi-objective training, enabling the generation of coherent summaries supported by informative and complementary visual content.

3 Methodology

As shown in Figure 2, `SPeCTrA-Sum` consists of two main stages: (a) representation and summarization, where the input text and images are encoded and fused to produce a summary conditioned on both modalities; and (b) image selection and alignment, where a subset of salient images is selected and aligned with the generated summary. The entire system is trained end-to-end using a multi-objective loss that jointly optimizes for summary quality, visual representativeness, and cross-modal semantic alignment.

3.1 Problem Setup and Base Architecture

A multimodal document consists of text $X = \{x_1, \ldots, x_{T_x}\}$ and images $\mathcal{I} = \{I_1, \ldots, I_M\}$. The objective is to generate an abstractive summary $Y = \{y_1, \ldots, y_{T_y}\}$ and a representative image subset $\mathcal{I}^\star \subseteq \mathcal{I}$ that reflects salient visual content aligned with the summary. Formally, the model learns a conditional mapping:

$$f_\theta : (X, \mathcal{I}) \longrightarrow (Y, \mathcal{I}^\star), \tag{1}$$

where $Y = f_\theta^{\text{text}}(X, \mathcal{I})$ is produced by the decoder, and $\mathcal{I}^\star = f_\theta^{\text{vis}}(\mathcal{I})$ is predicted by the VRP. Both outputs are trained jointly to ensure that $\mathcal{I}^\star$ is (i) semantically aligned with Y and (ii) diverse, promoting textual–visual coherence and visual complementarity.

Backbone Architecture. We adopt LLaVA-OneVision (Li et al., 2025a) as our multimodal scaffold, which uses Qwen-2 (Wang et al., 2024) as the causal language model (LLM) for generation. In this setup, a frozen vision encoder (SigLIP by Zhai et al. (2023)) is connected to the LLM via a multimodal projector, mapping visual features into the same token embedding space. This enables token-level concatenation of image and text inputs, a design widely used in LLaVA and MiniGPT-4 (Liu et al., 2023; Zhu et al., 2024).

Each visual vector is then projected into the LLM's token space via a learned linear mapping:

$$\tilde{\mathbf{v}}_t^{(m)} = P\left(\mathbf{v}_t^{(m)}\right). \tag{2}$$

After projection, features from both modalities are concatenated. This "projection-then-concatenation" strategy aligns vision and language tokens in a shared space.

Limitations of Pure Concatenation. Concatenating all projected visual tokens $\tilde{v}$ at the prefix of a decoder-only LLM introduces two major issues. First, in causal architectures, prefix tokens exert less influence in deeper layers due to attention bias toward recent tokens. Second, many visual tokens may be irrelevant or redundant, resulting in inefficient attention usage. Moreover, the raw projector outputs often exhibit a representation gap relative

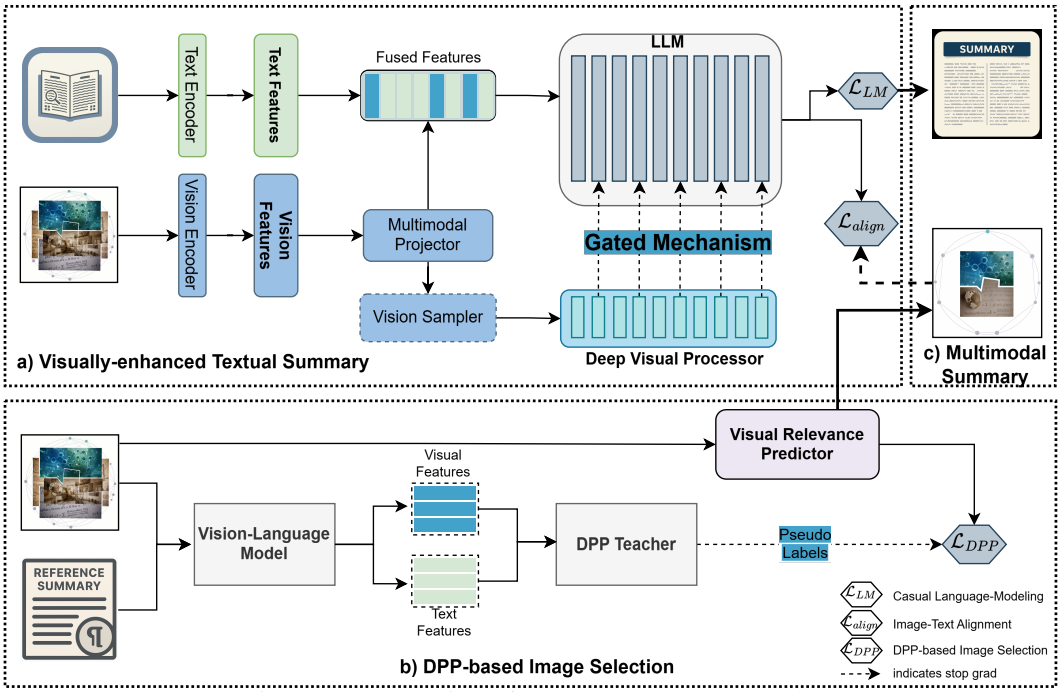


Figure 2: An overview of SPeCTrA-Sum Framework.

to the deeply transformed LLM hidden states, leading to poor cross-modal alignment.

These limitations motivate our proposed approach, which introduces hierarchically aligned and semantically filtered fusion mechanisms via the Deep Visual Processor (DVP) and VRP.

3.2 Proposed Method

We propose a unified framework that enhances multimodal summarization through five key components. Three modules are introduced to improve text–image fusion and representation alignment: (i) a vision sampler that filters out uninformative visual inputs, (ii) a DVP that enables layer-wise alignment between visual and textual features, and (iii) a Gated Integration mechanism that allows the model to dynamically control the contribution of visual features at different decoding depths. To further enable output-level image selection, we incorporate two additional components: (iv) a DPP-based Pseudo-Label Generator that produces soft inclusion probabilities encouraging both relevance and diversity, and (v) a lightweight image selector trained to learn from these labels, enabling efficient selection of salient, non-redundant images at inference time.

3.2.1 Vision Sampler

To reduce redundancy and retain only the most informative visual signals, we compress each image's patch grid into a fixed set of latent tokens using a Perceiver-style cross-attention bottleneck. Unlike top-K patch selection, this mechanism allows the model to learn what to retain via trainable latent queries.

Our design follows the Perceiver family of models (Jaegle et al., 2022), where learned latent vectors attend to the full input via cross-attention and are refined through self-attention. Flamingo (Alayrac et al., 2022) similarly uses a Perceiver Resampler to compress visual inputs into a fixed set of media tokens for efficient language integration.

Formally, for an image I_m represented by patch features $\mathbf{V}^{(m)} \in \mathbb{R}^{T_v \times d_v}$, and corresponding positional encodings $\mathbf{P}^{(m)} \in \mathbb{R}^{T_v \times d_v}$, we produce a compressed representation using a Perceiver stack of depth D. The vision sampler maintains a trainable latent array $\mathbf{L}^{(0)} \in \mathbb{R}^{L \times d}$, where L is the number of compressed visual tokens output per image. These latents serve as compact, semantically rich visual summaries, passed downstream to the language model via multimodal fusion.

3.2.2 Deep Visual Processor

To address the representation gap between shallow visual embeddings and deep language model activations, we introduce a DVP that refines visual features through a stack of transformer layers aligned with the LLM's depth.

Specifically, the compressed visual tokens produced by the Vision Sampler are processed through a sequence of L transformer blocks:

$$\hat{\mathbf{v}}^{(\ell)} = \mathrm{DVP}^{(\ell)}\left(\hat{\mathbf{v}}^{(\ell-1)}\right), \quad \ell = 1, \ldots, L, \quad (3)$$

where each DVP block consists of a self-attention layer followed by a feed-forward MLP, and matches the hidden dimension d of the corresponding LLM layer. This results in depth-aligned visual representations $\hat{\mathbf{v}}^{(\ell)}$, which evolve in parallel with the LLM's hidden states $\mathbf{v}^{(\ell)}$. This alignment enables layer-wise multimodal fusion, ensuring that visual information is semantically compatible with the abstraction level of each LLM layer.

3.2.3 Layer-Aligned Gated Cross-Attention

To enable controlled and depth-aware fusion of vision into the LLM, we insert gated cross-attention modules at every n-th layer in the decoder. At each injection point $\ell \in \{n, 2n, \ldots\}$, depth-aligned visual tokens $\hat{\mathbf{v}}^{(\ell)}$ attend to the LLM hidden states via a tanh-gated residual connection, inspired by Flamingo's cross-attention strategy (Alayrac et al., 2022).

Each gate is initialized near zero, allowing the model to initially preserve the behavior of the base LLM and gradually learn to integrate visual input. This mechanism supports progressive, learnable fusion aligned with the semantic hierarchy of the decoder. Full attention equations and gating details are provided in Appendix A.1.

3.2.4 Visual Relevance Predictor (VRP)

The VRP is a lightweight module that selects a subset of images that are both semantically relevant and mutually diverse, enabling the summarizer to condition on the most informative visual content. VRP is trained in a student–teacher distillation framework, where a principled DPP (Macchi, 1975; Kulesza and Taskar, 2012) acts as the teacher, producing soft selection probabilities that encode a trade-off between quality and diversity. A compact student network is trained to approximate these probabilities using only image embeddings, allowing for efficient, text-free inference that preserves the DPP's inductive biases.

DPP pseudo-labels. During training, the DPP acts as a teacher that assigns each image a soft inclusion probability $\pi_i \in [0, 1]$ based on its relevance to the summary and its redundancy with other images. These probabilities encode three key properties:

1. **Relevance**: Images closely aligned with the text receive higher probabilities.
2. **Diversity**: Redundant images are down-weighted through DPP structure.
3. **Cardinality control**: The expected set size $\sum_i \pi_i \approx \mu$ is enforced explicitly.

The DPP kernel is constructed using image-text relevance scores and an RBF-based (Lowe and Broomhead, 1988) diversity matrix over image embeddings. We refer readers to Appendix A.2 for full derivation of the kernel and marginal inclusion computation.

Student Network. The student network, VRP, is a small feed-forward predictor that maps each image embedding to a scalar selection logit. It operates independently on each image and does not require textual input at inference time. Given a normalized image embedding v_i, the predictor outputs:

$$z_i = f_{\mathrm{VRP}}(v_i), \quad (4)$$

where f_{VRP} is a two-layer MLP with GELU and dropout. The output p_i approximates the teacher's inclusion score π_i without requiring text. Since the DPP already captures set-level interactions, the student is trained to mimic these marginals independently for each image (see Section 3.3.3 for the image selection loss).

Inference Efficiency. While DPP-based selection requires expensive matrix operations and text conditioning ($\mathcal{O}(K^3)$, where $\mathcal{K}$ is the matrix dimension), the distilled VRP enables text-free, constant-time selection per image. This achieves practical inference speed while retaining DPP-style inductive biases, including relevance, diversity, and cardinality control. Empirically, the student achieves a strong approximation of the teacher and improves the quality of image subsets used for summarization.

3.3 Training Objective

The model is trained using a multi-objective (multimodal) loss that jointly optimizes three components:

1. an autoregressive language modeling loss for summary generation,
2. an image–text alignment loss to encourage cross-modal coherence, and
3. a distillation loss for training the VRP to mimic the DPP teacher.

Formally, the loss is defined as:

$$\mathcal{L}_{\text{MM}} = \underbrace{\mathcal{L}_{\text{LM}}}_{\text{causal LM}} + \lambda_{\text{align}} \underbrace{\mathcal{L}_{\text{align}}}_{\text{image–text alignment}} + \lambda_{\text{VRP}} \underbrace{\mathcal{L}_{\text{DPP}}}_{\text{image selection}}, \quad (5)$$

where λ_{align} and λ_{VRP} are hyperparameters that weight the auxiliary objectives relative to the main summarization loss.

3.3.1 Autoregressive summarization

Given a target token sequence $Y = \{y_1, \ldots, y_T\}$, source text x, and a corresponding set of images $\mathcal{I}$, the model is trained using teacher forcing to maximize the likelihood of the correct summary tokens. The autoregressive language modeling loss is defined as:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, x, \{\mathcal{I}\}), \quad (6)$$

where p_θ denotes the model's conditional probability distribution over tokens, given the multimodal input and previously generated tokens.

3.3.2 Image-Text Alignment

To ensure that generated summaries are semantically consistent with selected images, we introduce a global alignment objective between a projected decoder representation and a frozen visual embedding. Specifically, we align the decoder's mean-pooled hidden state to the average SigLIP embedding of the selected images, using a SigLIP-style contrastive loss.

The alignment encourages the decoder to ground textual output in visual semantics, while penalizing mismatches with unrelated image sets. Full formulation, including projection and contrastive loss computation, is provided in Appendix A.3.

3.3.3 Image Selection via DPP Distillation

To learn image selection that is both relevant and diverse, we train the VRP module to approximate soft inclusion probabilities generated by a DPP teacher. These probabilities encode a principled balance between text-image relevance, visual redundancy, and set cardinality.

During training, we compute the teacher probabilities using a DPP kernel constructed from text–image similarity and inter-image distances in a frozen SigLIP embedding space. The VRP then learns to match these soft labels using a calibrated cross-entropy loss, along with a regularization term to enforce target subset size μ.

This distillation allows the VRP to internalize the DPP's inductive biases while enabling fast, text-agnostic image selection at inference. Full derivations and kernel construction are provided in Appendix A.4.

4 Results

4.1 Implementation Details

We trained our model using a batch size of 1 with the Adafactor optimizer, following Marek et al. (2025), who showed that smaller batch sizes can be more effective than larger ones in certain settings. Training was controlled by step count, where one epoch corresponds to approximately 295k steps. Different systems were trained up to 360k steps, and the best model was selected based on validation loss. Further technical details are provided in Appendix A.5 and Appendix A.6.[1]

4.2 Comparison with Existing Systems

Besides the proposed DVP, we evaluate two variants that are simplifications of full DVP system: (i) **OneVision**, the base LLaVA-OneVision model; (ii) **Vision Sampler**, which omits the DVP and directly feeds visual features into the gating mechanism; and (iii) **DVP**, the full model shown in Figure 2.

Since existing automatic benchmarks are primarily optimized for textual overlap, we report ROUGE to contextualize textual quality, while also reporting the standard MSMO-style visual metrics available from prior work. For our models we additionally report image-text alignment and within-set diversity proxies, noting that many published baselines do not provide these values, which limits direct comparisons on those axes.

Table 1 presents a comparison between our proposed model and a range of published multimodal summarization systems evaluated on the MSMO benchmark. These systems include early baselines like ATG, ATL, and HAN (Zhu et al., 2018), as well as more recent models such as MOF (Zhu et al., 2020), UniMS (Zhang et al., 2022), SITA (Jiang et al., 2023), ViL-Sum (Cui et al., 2024), and DIUSum (Xiao et al., 2024).

Among all systems, ViL-Sum achieves the highest ROUGE scores (44.29 for ROUGE-1 and 20.96 for ROUGE-2), indicating strong performance on textual overlap. SITA, on the other hand, achieves

[1]Code: https://github.com/abidmeeraj/SPeCTrA-Sum

	Model	ROUGE-1	ROUGE-2	IP	MaxSim	MMAE
Baselines	ATG (2018)	40.63	18.12	59.28	25.82	3.35
	ATL (2018)	40.86	18.27	62.44	13.26	3.26
	HAN (2018)	40.82	18.30	61.83	12.22	3.25
	MOF (2020)	41.20	18.33	65.45	26.38	3.37
	UniMS (2022)	42.94	20.50	69.38	29.72	-
	SITA (2023)	43.64	20.53	76.41	33.47	3.37
	BART-VGG (2024)	43.75	20.70	-	-	-
	ViL-Sum (2024)	44.29	20.96	66.27	32.17	3.55
	DIUSum (2024)	42.23	19.83	-	-	-
	BERTAbs (2024)	41.85	19.40	-	-	-
Proposed Model	DVP (ours)	44.20	20.77	74.03	31.68	3.55

Table 1: Results of published baseline models in existing studies.

System	R-1	R-2	BERTScore	IP	CLIPScore	MMAE	PCD
OneVision	43.81	20.52	89.58	74.02	70.62	3.5447	32.66
Vision Sampler	44.06	20.78	89.53	74.01	70.54	3.5484	32.65
DVP	**44.20**	20.77	89.33	**74.03**	70.52	**3.5521**	**32.81**

Table 2: Model performance under multi-objective loss ($\mathcal{L}_{\text{MM}}$). PCD = Pairwise Cosine Distance (diversity).

the highest Image Precision (IP) at 76.41, reflecting more accurate image selection with respect to ground-truth annotations.

Our model (DVP), trained with a multi-objective loss, achieves ROUGE-1/44.20 and ROUGE-2/20.77, effectively matching ViL-Sum in text generation performance (a gap of only 0.09 and 0.19, respectively). In terms of image selection quality, DVP attains IP/74.03, surpassing ViL-Sum (66.27) and approaching SITA's upper bound, while maintaining strong MaxSim and MMAE scores. This reflects DVP's ability to select images that are both relevant and semantically aligned with the summary.

These results are significant for two reasons. First, they confirm that our model competes effectively on traditional metrics despite incorporating deeper visual processing, even though such metrics are not well-suited for evaluating systems with stronger visual grounding. Second, they validate that DVP's image selection is not only quantitatively strong but also better aligned with the multimodal summarization objective, balancing informativeness, relevance, and non-redundancy more effectively than many existing approaches.

4.3 Impact of Multi-Objective Training

Table 2 shows that multi-objective training improves both textual and visual quality. DVP recovers +0.39 ROUGE-1 and +0.19 ROUGE-2 compared to its single-objective counterpart (Table 3), while achieving the best visual diversity and IP among all evaluated models.

System	R-1	R-2	BERTScore
OneVision	**44.26**	**20.86**	89.12
Vision Sampler	43.89	20.61	**89.54**
DVP	43.81	20.58	89.50

Table 3: Performance under MaskedLM training.

4.4 Effect of Visual Processing Under MaskedLM

Table 3 isolates the effect of introducing stronger visual processing under a MaskedLM training objective. Moving from OneVision to Vision Sampler and then to DVP reduces ROUGE (from 44.26/20.86 for OneVision to 43.81/20.58 for DVP), although BERTScore remains broadly comparable (89.12–89.54). This indicates that deeper visual processing does not automatically improve text-centric metrics and can even degrade n-gram overlap under a purely textual objective. This observation motivates our multi-objective formulation, which balances textual and visual objectives, partially recovers ROUGE, and improves visual selection metrics (Table 1).

5 Analysis

5.1 Human Evaluation

Automatic metrics alone may fail to capture fine-grained aspects of multimodal quality, particularly the alignment between textual and visual content. To address this limitation, we conduct a human evaluation of our DVP-based system.

We sample 200 articles and collect three independent annotations per article via Amazon Mechanical Turk, resulting in 600 valid responses from 507

Dimension	Mean (SD)	% ≥4	Exact	W1
Text quality	3.90 (0.69)	80.1	49.0	90.0
Image relevance	4.04 (0.80)	76.8	44.3	84.0
Image diversity	3.89 (0.83)	73.2	43.0	82.2
Overall quality	4.00 (0.71)	79.2	45.8	85.5

Table 4: Human evaluation results. We report mean ratings and standard deviation (SD), the proportion of ratings ≥ 4, and inter-annotator agreement by reporting exact match and within-one agreement (W1) percentage.

System	w Rel. Filter	w/o Rel. Filter
	Mean / Max	Mean / Max
OV	32.66 / 36.53	36.72 / 43.19
Vision Sampler	32.65 / 36.50	36.73 / 43.22
DVP	**32.81 / 36.71**	**36.77 / 43.27**

Table 5: Pairwise cosine distance among selected images (higher = more diverse).

unique workers after filtering 2.1% of submissions using attention checks. Annotators rate each example on a 5-point Likert scale across four dimensions: text quality, image relevance, image diversity, and overall multimodal quality.

As shown in Table 4, the system achieves consistently strong ratings across all dimensions, with most responses in the high-quality range ($\geq$4). Image relevance attains the highest average score, indicating strong alignment between images and text. Text quality and overall multimodal quality are also rated highly, suggesting that improved visual grounding does not compromise textual coherence.

Although image diversity scores are slightly lower, they remain positive, reflecting a reasonable balance between relevance and variety. Inter-annotator agreement is high for a subjective task, particularly under the within-one criterion, indicating consistent preferences among annotators. Overall, these findings provide complementary human evidence that the system produces coherent, visually grounded summaries.

5.2 Relevance vs. Diversity Trade-off

As shown in Table 5, diversity scores (measured via pairwise cosine distance) are highly sensitive to relevance filtering. Without filtering, semantically unrelated images inflate diversity scores artificially. After thresholding, the scores become more meaningful: DVP maintains the highest mean and max diversity among filtered images, demonstrating its ability to select both relevant and non-redundant content.

While embedding-based thresholds are effective for filtering irrelevant images, our qualitative analysis (Figure 3) reveals that some discarded images still provide useful contextual cues, such as related events or background details. Although not directly aligned with the summary content, these images can enrich reader understanding and serve a complementary role rather than being strictly irrelevant.

This raises an open question: How can models distinguish between irrelevant content and visually complementary information? Relying solely on hard similarity thresholds may discard valuable context. Future work should explore modeling complementarity alongside relevance and diversity, possibly incorporating user-centric or task-aware signals to better assess an image's supportive value.

5.3 Qualitative Analysis of Visual Grounding

Our qualitative analysis on generated summaries (Figures 4–5) reveals a clear progression in model awareness. While baseline models (OV, Vision Sampler) exhibit a "textual lead bias" focused on procedural facts, DVP leverages visual saliency as a priority filter. DVP successfully extracts buried descriptive details, such as garment textures, specific accessories, and cultural markers, that are visually prominent in the source images but ignored by text-centric models. This demonstrates DVP's unique ability to re-prioritize information through cross-modal grounding, yielding summaries that better reflect the human viewing experience.

6 Conclusion

We introduced `SPeCTrA-Sum`, a multimodal summarization framework that elevates the role of visual evidence in text generation and performs principled, efficient image subset selection. Our approach incorporates a Deep Visual Processor (DVP) for layer-aligned visual conditioning and a lightweight Visual Relevance Predictor (VRP) distilled from a diversity-aware DPP teacher. Together with multi-objective training, this design allows the model to produce not only fluent summaries but also visually grounded outputs that highlight the most informative elements of the input. Overall, `SPeCTrA-Sum` offers a cohesive and efficient solution to multimodal summarization, balancing informativeness, fluency, and visual complementarity.

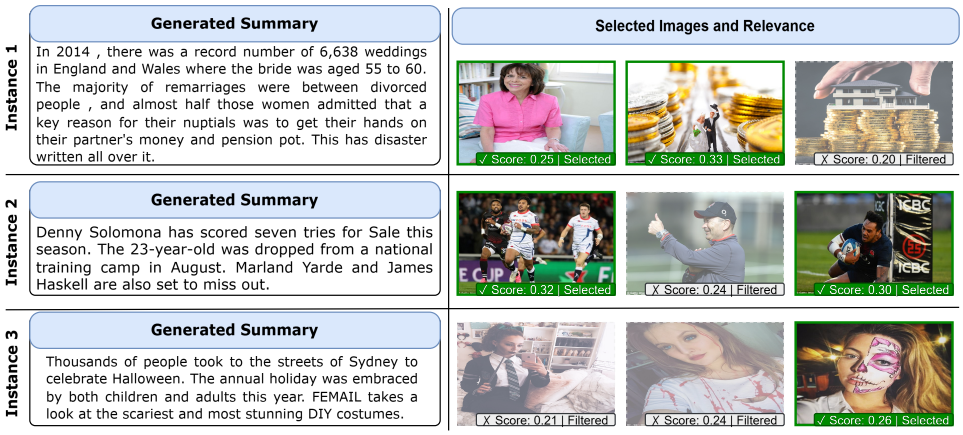


Figure 3: Examples from the test set illustrating image–summary relevance. For this experiment, a similarity threshold of 0.25 was used to filter out low-relevance images..

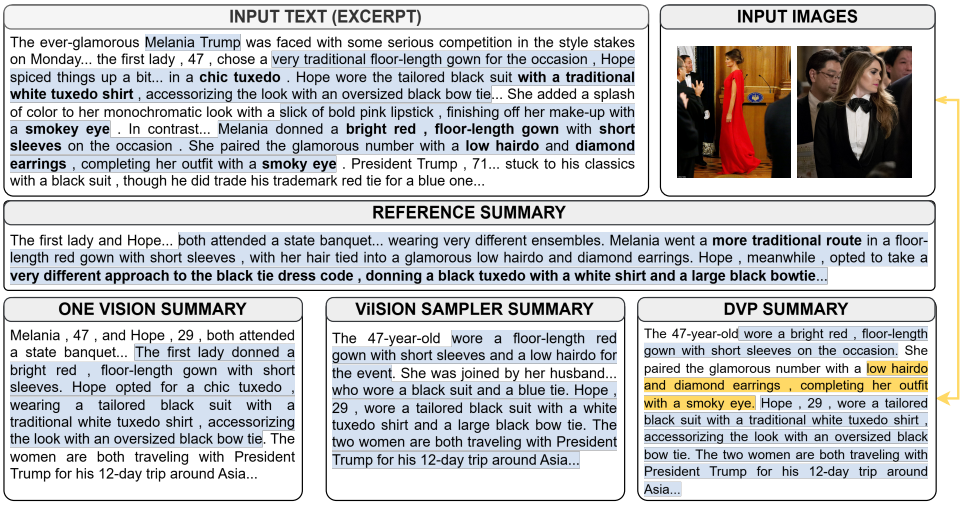


Figure 4: Comparison of summaries for a state banquet example. OV and Vision Sampler capture general descriptions (blue), whereas DVP includes fine-grained visual details (yellow), such as “diamond earrings” and “smoky eye”.

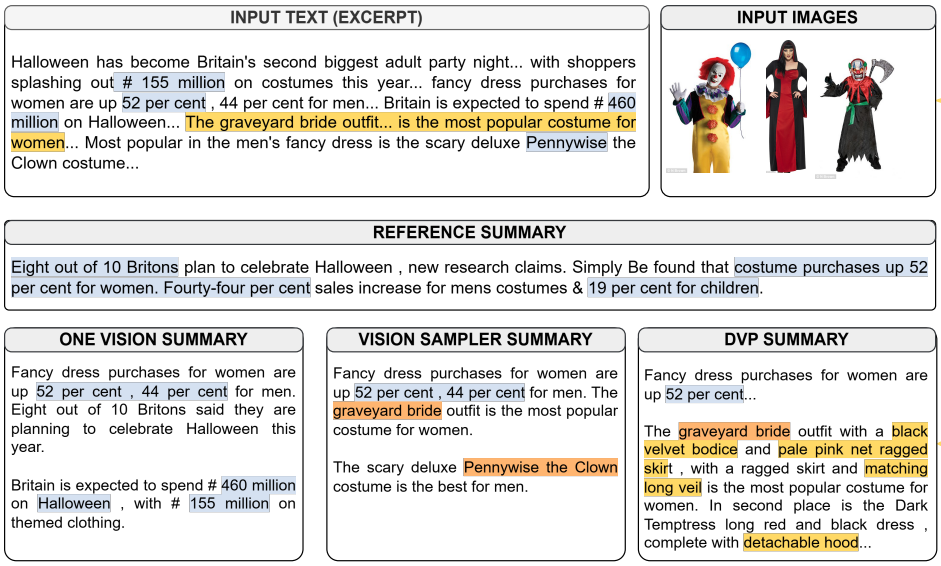


Figure 5: Comparison of summaries for Halloween costume trends. OV captures text-based statistics (blue), Vision Sampler includes basic entities (orange), and DVP adds fine-grained visual details (yellow) absent from the text.

Limitations

Despite promising results, our study highlights several limitations in current evaluation practices and modeling assumptions for multimodal summarization.

First, standard automatic metrics like ROUGE remain poorly aligned with the goals of visually grounded generation. They measure n-gram overlap but not whether selected images meaningfully supplement the summary. While we report MSMO-style visual metrics and used additional proxies (e.g., CLIPScore, pairwise cosine distance), prior systems do not report these, limiting direct comparability.

Second, As discussed in Section 5.2, diversity scores can be inflated by semantically irrelevant images, complicating interpretation unless standardized relevance filtering is applied.

Finally, while our evaluation focuses on automatic metrics, incorporating human judgments of visual-textual grounding, complementarity, and user-perceived utility would offer valuable additional insights. Developing such assessments and benchmarks could help more comprehensively capture multimodal coherence and further validate system performance.

Ethical Considerations

Our work uses the MSMO dataset, derived from public news sources, which may reflect inherent social or cultural biases. While our model improves visual-textual alignment, it may inherit such biases and reflect subjective notions of image relevance. Future work should explore fairness-aware training and human evaluation to ensure broader and more responsible applicability.

Acknowledgments

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Kumbhar Atharva, Kulkarni Harsh, Mali Atmaja, Sonawane Sheetal, and Mulay Prathamesh. 2023. The current landscape of multimodal summarization. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 797–806.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

J Cao, Z Gan, Y Cheng, L Yu, YC Chen, and J Liu. 2020. Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models. In *Computer Vision – ECCV 2020*, pages 565–580, Cham. Springer International Publishing.

L Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. 2018. Fair and diverse DPP-based data summarization. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 716–25.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. 2025. Multimodal Language Models See Better When They Look Shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6678–6695.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive Text-Image Summarization Using Multi-Modal Attentional Hierarchical RNN. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4046–4056.

Sangwoo Cho, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2019. Multi-Document Summarization with Determinantal Point Processes and Contextualized Representations. pages 98–103.

Chenhao Cui, Xinnian Liang, Shuangzhi Wu, and Zhoujun Li. 2024. Align vision-language semantics by multi-task learning for multi-modal summarization. *Neural Computing and Applications*, 36(25):15653–15666.

S Gupta, J Hoffman, and M Jitendra. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Andrew Jaegle, Sebastian Borgeaud, Jean Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. 2022. PERCEIVER IO: A GENERAL ARCHITECTURE FOR STRUCTURED INPUTS & OUTPUTS. *ICLR 2022 - 10th International Conference on Learning Representations*.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A survey on multi-modal summarization. *ACM Computing Surveys*, 55(13s):1–36.

Chaoya Jiang, Rui Xie, Wei Ye, Jinan Sun, and Shikun Zhang. 2023. Exploiting Pseudo Image Captions for Multimodal Summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 161–175. Association for Computational Linguistics (ACL).

A Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *nowpublishers.com*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*, 2025.

H Li, J Zhu, T Liu, and J Zhang. 2018. Multi-modal Sentence Summarization with Modality Attention and Image Filtering. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

Junnan Li, Ramprasaath R Selvaraju, Akhilesh D Gotmare, Shafiq Joty, Caiming Xiong, and Steven C H Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. 2025b. STI-Bench: Are MLLMs Ready for Precise Spatial-Temporal World Understanding? *arXiv preprint arXiv:2503.23765*.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2023. Summary-Oriented Vision Modeling for Multimodal Abstractive Summarization. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:2934–2951.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *Advances in neural information processing systems*, 36:34892–34916.

X Liu and H Liu. 2025. Symmetry-Aware Advances in Multimodal Large Language Models: Architectures, Training, and Evaluation. *MDPI*, 17(9).

D Lowe and D Broomhead. 1988. Multivariable functional interpolation and adaptive networks. *Complex systems*, pages 321–355.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA.

Odile Macchi. 1975. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122.

Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Gordon Wilson, and Micah Goldblum. 2025. Small Batch Size Training for Language Models: When Vanilla SGD Works, and Why Gradient Accumulation Is Wasteful.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, Youliang Yan, Huawei Noah', and Ark Lab. 2023. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18983–18992.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:1073–1083.

Hao Tan and Mohit Bansal. 2019. LXMert: Learning cross-modality encoder representations from transformers. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 5100–5111.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution.

Min Xiao, Junnan Zhu, Haitao Lin, Yu Zhou, and Chengqing Zong. 2023. CFSum: A Coarse-to-Fine Contribution Network for Multimodal Summarization. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:8538–8553.

Min Xiao, Junnan Zhu, Feifei Zhai, Yu Zhou, and Chengqing Zong. 2024. DIUSum: dynamic image utilization for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19297–19305.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. 2024. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15952–15962.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022. UniMS: A unified framework for multimodal summarization with knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11757–11764.

Zhicheng Zhang, Yibo Sun, and Shiyan Su. 2023. Multimodal Learning for Automatic Summarization: A Survey. In *International Conference on Advanced Data Mining and Applications*, pages 362–376.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MINIGPT-4: ENHANCING VISION-LANGUAGE UNDERSTANDING WITH ADVANCED LARGE LANGUAGE MODELS. *12th International Conference on Learning Representations, ICLR 2024*.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164, Brussels, Belgium. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9749–9756.

A Appendix

A.1 Gated Cross-Attention Injection in Decoder

This appendix provides full implementation details for the gated cross-attention mechanism used to inject visual features into the decoder. We include the formulation of layer-wise cross-attention, gating behavior, and integration within the transformer block stack.

Let $\mathbf{h}^{(\ell)} \in \mathbb{R}^{T\times d}$ denote the LLM hidden states at layer ℓ, and let the keys and values be derived from the depth-aligned visual states $\mathbf{v}^{(\ell)}$. The cross-attention output is computed as:

$$\mathrm{XAttn}^{(\ell)}(\mathbf{h}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}) = \mathrm{softmax}\left(\frac{Q^{(\ell)}\mathbf{h}^{(\ell)}\left(K^{(\ell)}\hat{\mathbf{v}}^{(\ell)}\right)^{\top}}{\sqrt{d_k}}\right) V^{(\ell)}\hat{\mathbf{v}}^{(\ell)}, \tag{7}$$

where $Q^{(\ell)}, K^{(\ell)}, V^{(\ell)}$ are learned projection matrices for queries, keys, and values respectively.

To modulate the contribution of the visual signal, we introduce a trainable gate $\alpha^{(\ell)} \in \mathbb{R}$, initialized close to zero. The residual connection is updated as:

$$\tilde{\mathbf{h}}^{(\ell)} = \mathbf{h}^{(\ell)} + \tanh\left(\alpha^{(\ell)}\right) \odot \mathrm{XAttn}^{(\ell)}(\mathbf{h}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}), \tag{8}$$

where $\odot$ denotes element-wise multiplication. Because $\tanh\left(\alpha^{(\ell)}\right) \approx 0$ at initialization, the model initially behaves like the original LLM, and learns over time how much visual information to integrate at each injection layer.

Between injection points, the LLM proceeds as usual with standard transformer blocks:

$$\mathbf{h}^{(\ell+1)} = \mathrm{LLMBlock}^{(\ell)}\left(\tilde{\mathbf{h}}^{(\ell)}\right). \tag{9}$$

This design ensures that visual features are progressively and selectively fused, respecting both the depth and semantic abstraction level of the LLM's internal representations.

A.2 DPP Teacher Kernel and Marginal Computation

The DPP kernel is constructed as:

$$L = Q^{1/2}\kappa Q^{1/2} + \varepsilon I, \tag{10}$$

where $Q = \mathrm{diag}(q_1, \ldots, q_K)$ encodes per-image relevance scores, κ is an RBF-based similarity matrix capturing inter-image redundancy, and εI ensures numerical stability. To control the expected subset size, we define the marginal

kernel $L(L + tI)^{-1}$, and solve for t^* such that $\mathrm{Tr}(K(t^*)) = \mu$, the desired cardinality.

The resulting marginal inclusion probabilities are:

$$\pi_i = \frac{\lambda_i}{\lambda_i + t^*}, \tag{11}$$

where λ_i are the eigenvalues of L.

A.3 Global Image-Text Alignment Objective

We encourage semantic consistency between the decoder's representation and the selected images using a SigLIP-inspired global contrastive loss.

Let $\mathrm{SigText}$ and $\mathrm{SigImage}$ be frozen encoders that map inputs into ℓ_2-normalized vectors in $\mathbb{R}^{d_p}$. For an article a, with reference summary Y_a and selected images $\{I^a_{i_k}\}_{k=1}^K$, the teacher embeddings are defined as:

$$\begin{aligned} t_{\text{sig}}(a) &= \mathrm{norm}\big(\mathrm{SigText}(Y_a)\big), \\ i^a_k &= \mathrm{norm}\big(\mathrm{SigImage}(I^a_{i_k})\big), \\ v_{\text{sig}}(a) &= \mathrm{norm}\Big(\tfrac{1}{K}\sum_{k=1}^{K} i^a_k\Big). \end{aligned} \tag{12}$$

Here, $v_{\text{sig}}(a)$ is the mean pooled image embedding for article a, encoding the visual summary.

On the student side, let $H^{\text{dec}} \in \mathbb{R}^{T\times d}$ denote the decoder hidden states under teacher forcing. We compute the student's pooled representation via:

$$s = \mathrm{meanpool}(H^{\text{dec}}),\ t_{\text{stu}} = \mathrm{norm}\big(g_{\text{text}}(s)\big)$$

where $g_{\text{text}} : \mathbb{R}^d \to \mathbb{R}^{d_p}$ is a learnable projection module that maps decoder features to the alignment space.

We define a logistic similarity function:

$$z(x, y) = \tfrac{1}{\tau}\, x^\top y \tag{13}$$

with temperature parameter $\tau > 0$, and draw a set of negatives $\mathcal{S}$ from other articles in the batch or dataset.

The image-text alignment loss is then formulated in a SigLIP-style contrastive objective:

$$\begin{aligned} \mathcal{L}_{\text{align}} = &-\log \sigma\big(z(t_{\text{stu}},\, v_{\text{sig}}(a))\big) \\ &- \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \log\big(1 - \sigma(z(t_{\text{stu}},\, v_{\text{sig}}(j)))\big) \\ &- \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \log\big(1 - \sigma(z(t_{\text{sig}}(j),\, v_{\text{sig}}(a)))\big). \end{aligned} \tag{14}$$

where σ is the sigmoid function. The first term maximizes similarity between the student text and corresponding visual summary. The second penalizes similarity between the student text and mismatched visual contexts. The third penalizes mismatched teacher text–image pairs to regularize the frozen teacher space.

This alignment objective encourages the student decoder to produce summaries that are semantically grounded in the selected images, while also aligning to a frozen, high-quality multimodal embedding space.

A.4 DDP-Based Image Selection and Distillation

This appendix provides full details on the construction of the DPP teacher used to supervise the VRP, including the derivation of marginal probabilities and the training objective used for distillation.

A.4.1 DPP Teacher Construction

Let $I^{aK}_{i\ i=1}$ be the candidate images for a document or article a, with reference summary Y_a. The DPP teacher produces a probability distribution over subsets of images that favors both high relevance to the text and inter-image diversity.

We use a frozen SigLIP model to extract ℓ_2-normalized embeddings for the summary and each image:

$$\begin{aligned} e_{\text{text}} &= \mathrm{norm}\big(\mathrm{SigText}(Y_a)\big), \\ e_i &= \mathrm{norm}\big(\mathrm{SigImage}(I^a_i)\big). \end{aligned} \tag{15}$$

The relevance score for image i is computed as the inner product $r_i = e^\top_{\text{text}} e_i$, and similarity between image i and j as $s_{ij} = e_i^\top e_j$.

We compute per-image quality scores as:

$$q_i = \exp(\gamma r_i)$$

where $\gamma > 0$ is a scaling factor.

To model diversity, we define an RBF kernel on the cosine distances between image embeddings:

$$D^2_{ij} = 2\,(1 - s_{ij}),\ \kappa_{ij} = \exp\Big(-\tfrac{D^2_{ij}}{2\sigma^2}\Big),\ \kappa_{ii} = 1,$$

DPP Kernel: We construct a quality-diversity kernel $L \in \mathbb{R}^{K\times K}$ as:

$$L\ =\ Q^{1/2}\,\kappa\,Q^{1/2}\ +\ \varepsilon I,\ Q = \mathrm{diag}(q_1, \ldots, q_K), \tag{16}$$

where $\varepsilon > 0$ is a small constant for numerical stability. This kernel defines a DPP over image subsets.

A.4.2 Marginal Inclusion Probabilities

To control the expected number of selected images μ, we define the marginal kernel:

$$K(t) \;=\; L\,(L+tI)^{-1}, \quad t \geq 0, \tag{17}$$

and solve for t^* such that:

$$\operatorname{Tr} K(t^*) = \mu$$

Using eigendecomposition $L = U\,\operatorname{diag}(\lambda_\ell)\,U^\top$, we compute:

$$\begin{aligned} \operatorname{Tr} K(t) &= \sum_\ell \frac{\lambda_\ell}{\lambda_\ell + t}, \\ K(t) &= U\,\operatorname{diag}\!\Big(\tfrac{\lambda_\ell}{\lambda_\ell+t}\Big)\,U^\top. \end{aligned} \tag{18}$$

The final soft marginal inclusion probabilities used for distillation are:

$$\pi_i \;=\; [K(t^*)]_{ii} \;\in\; [0,1], \quad \sum_{i=1}^{K} \pi_i \;\approx\; \mu, \tag{19}$$

These soft labels encoder: (1) Relevance (via q_i), (2) Diversity (via κ), and (3) Cardinality control (via μ constraint).

A.4.3 VRP Student Objective

Let $z_i = f_{\text{VRP}}(v_i)$ denote the scalar logit predicted by the VRP for image i, and $p_i = \sigma(z_i)$ the corresponding selection probability.

We train the VRP to match the DPP teacher marginals using a stable logit-level binary cross-entropy, plus a soft constraint on expected cardinality:

$$\begin{aligned} \mathcal{L}_{\text{VRP}} = & \underbrace{\tfrac{1}{K}\sum_{i=1}^{K}\Big(\operatorname{softplus}(z_i) - \pi_i z_i\Big)}_{\text{probability matching}} \\ & + \underbrace{\alpha\Big(\sum_{i=1}^{K} p_i - \mu\Big)^2}_{\text{cardinality regularization}}. \end{aligned} \tag{20}$$

where $\alpha \geq 0$ controls the strength of the regularization.

This distillation approach transfers the inductive bias of DPPs into a lightweight, text-agnostic image selector. It enables:

- $\mathcal{O}(K)$ inference (vs. $\mathcal{O}(K^3)$ for DPP sampling),
- Soft, diverse image selection at test time,
- And stable training via smooth supervision.

This loss complements our summarization and alignment objectives, resulting in coherent summaries grounded in relevant and diverse visual evidence.

A.5 Technical Implementation Details

A.5.1 Technical Details

We used LLaVA-OneVision with Qwen2-7B variant as our base model. On top of the base model, we introduced all the modules discussed in Section 3.2. Key training hyperparameters are listed in Table 6.

Hyperparameter	Value
Optimizer	Adafactor
Mixed Precision	bfloat16
Precision	bf16
Max input length	2048
Max images	5

Table 6: Key training hyperparameters.

We use a two-stage training procedure guided by the validation loss focusing on Architecture (Vision Sampler/DVP capacity and the set of gated decoder layers). Then LoRA fine-tuning parameters were optimized using the best architecture from stage 1. Search space and relevant hyperparameter are listed in Table 8.

Regarding image selection using VRP and DPP, the hyperparameter details and their interpretation is given in Table 7. Moreover, all experiments were conducted using a single NVIDIA A100 GPU with 80 GB memory. For memory efficiency, 4-bit QLoRA-style quantization technique is used.

A.5.2 Dataset Details

Sourced from the Daily Mail website, MSMO pairs each news article with all in-page images, and constructs reference summaries by combining article headlines. Human annotators also mark the images deemed most representative of the story. On average, each article contains approximately 650 words and nine images, offering a realistic and visually rich setting for multimodal summarization. The dataset provides both textual references (for ROUGE-style evaluation) and image relevance labels. Owing to its scale, permissive licensing,

H.parameter	Interpretation	Value
K	max selected images per example	3
σ	similarity kernel bandwidth (RBF)	0.8
γ	relevance scaling for q_i (quality weighting)	2.0
μ	target/expected subset size (cardinality prior)	3.0
α	strength of subset-size regularization	0.3
ϵ	numerical stability term for kernel matrix	1×10^{-5}
γ	relevance scaling for q_i (quality weighting)	2.0

Table 7: VRP-related hyperparameters.

and established leaderboard, MSMO has become the de-facto benchmark—often described as the "MNIST of multimodal summarization".

A.6 Computational Overhead Analysis

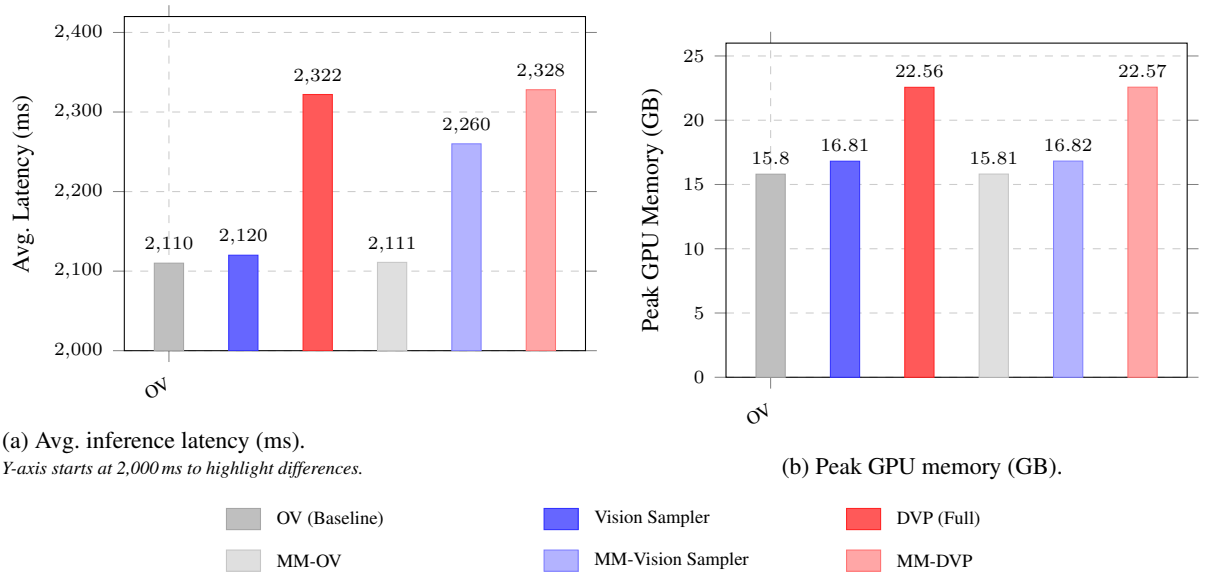
We provide a detailed computational overhead analysis comparing DVP against the simple concatenation baseline (OV) across all six model variants. All measurements were performed on a single NVIDIA A100-SXM4-80GB GPU using 100 test samples with 10 warm-up iterations, generating up to 256 tokens per sample under identical hardware and software conditions. Table 9 reports exact figures for all variants; Figure 6 visualises the absolute latency and memory cost side by side.

Group	Hyperparameter	Search Space and Optimized Values
Vision Sampler	Number of latents	[**32**, 64]
	Depth	[2, **4**]
	FF multiplier	[2, **4**]
DVP	Number of layers	[16, 20, **24**]
Gated Connections	Layer indices	[8, 16] with 16 DVP layers
		[8, 16] with 20 DVP layers
		[8, 16, 24] with 24 DVP layers
		[4, 8, ... , 24] with 24 DVP layers
LoRA	Rank r	[16, **32**, 64]
	α	[64, **128**]

Table 8: Search space for each hyperparameter group with optimal values highlighted in **bold**.

Variant	Architecture	Avg. Latency (ms)	Latency Overhead	Peak GPU Mem. (GB)	Memory Overhead
OV (Baseline)	Concatenation	∼2,110	—	15.80	—
Vision Sampler	Perceiver Only	2,120	+0.5%	16.81	+6.4%
DVP	DVP (Full)	2,322	+10.0%	22.56	+42.8%
MM-OV	Concat + MM Loss	2,111	+0.0%	15.81	+0.1%
MM-Vision Sampler	Perceiver + MM Loss	2,260	+7.1%	16.82	+6.5%
MM-DVP	DVP + MM Loss	2,328	+10.3%	22.57	+42.8%

Table 9: Inference speed and peak GPU memory across all six model variants. Measurements on a single NVIDIA A100-SXM4-80GB GPU (100 test samples, 10 warm-up iterations, up to 256 generated tokens).



(a) Avg. inference latency (ms).
Y-axis starts at 2,000 ms to highlight differences.

(b) Peak GPU memory (GB).

Figure 6: Absolute inference cost of all six model variants (exact figures in Table 9). **Left:** average latency per sample (ms); y-axis starts at 2,000 ms to accentuate differences — actual values are printed above each bar. **Right:** peak GPU memory (GB). Darker shades = base architecture; lighter shades = corresponding MM-loss variant (inference-identical to their base counterpart).