AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3040

ADAPTIVE CROSS-MODAL ALIGNMENT AND ORTHOGONAL TASK DECOUPLING FOR MULTITASK MULTIMODAL SUMMARIZATION

ZhiAng Chen
School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China.

Abstract: Producing a reliable multimodal summary requires a model to determine which claims are supported across language and vision and then express that evidence coherently. Training only for generation leaves useful cross-modal supervision untapped, yet indiscriminate parameter sharing across auxiliary tasks can merge transferable semantics with task-dependent cues and thereby induce negative transfer. Alignment is made still harder by visually salient background regions that have little bearing on the desired summary. To address these difficulties, we develop a single framework for multimodal summarization, image-text matching, and image-text retrieval. Visual tokens are first weighted through text-conditioned semantic attention; a learned gate then suppresses unhelpful visual responses before fusion. The aligned sequence is decomposed into one common representation and three private, task-oriented representations. Their redundancy is controlled by a Frobenius-norm orthogonality penalty, optimized jointly with the generation, matching, and retrieval objectives. The implementation couples a T5-base language encoder with CLIP ViT-B/16 and maintains 768-dimensional features in both streams. On MSMO, the full system records ROUGE-1, ROUGE-2, ROUGE-L, and BLEU values of 46.95, 22.03, 44.42, and 18.71, respectively, exceeding the single-task and multitask comparators included in the reported evaluation. Removing adaptive alignment, the shared-private decomposition, or the orthogonal term lowers ROUGE-L by 2.91, 2.08, and 0.87 points. Taken together, the findings indicate that text-directed visual filtering and an explicit division between reusable and task-specific information are both consequential for grounded summary generation.
Keywords: Multimodal summarization; Multitask learning; Orthogonal representation decoupling; Cross-modal alignment; Shared-private representation; Negative transfer

1 INTRODUCTION

A multimodal summary compresses evidence distributed across two or more modalities into a short textual account. In image-supported news, the article establishes the narrative, whereas the accompanying photograph may reveal entities, settings, attributes, or contextual details that the prose leaves unstated. Successful generation therefore involves more than shortening text: the system must identify useful visual observations, connect them to linguistic meaning, and verbalize only evidence that remains fluent and well grounded. MSMO captures this setting through news stories paired with images and human-authored summaries [1].

Many existing approaches either learn solely from the summarization loss or pass concatenated language and image features directly to a decoder. Although such models can discover cross-modal associations, sequence-generation supervision alone provides a narrow learning signal. Image-text matching asks a different question—whether two inputs convey compatible meaning—while retrieval requires the representation space to separate valid pairs from mismatched candidates. When learned together, these objectives can reinforce cross-modal agreement at both the pair and embedding levels instead of optimizing only the local statistics of the next output token.

Benefits from multitask training cannot be assumed. Hard sharing forces objectives with different demands to rely on the same intermediate coordinates. Summary generation prioritizes word choice and discourse continuity; matching operates on pairwise consistency; retrieval depends on fine discrimination among candidates. Updates from these branches may consequently compete. Once common information and task-bound cues become entangled, an improvement for one loss can erase a feature needed elsewhere—a familiar form of negative transfer. A shared-private factorization provides a more deliberate arrangement: broadly useful cross-modal meaning is assigned to a common subspace, while private projections retain the statistics needed by each task [2-3].

Interference also arises within the visual input itself. A photograph includes foreground actors, incidental objects, texture, and background scenery, much of which may never be mentioned in the article. Unfiltered fusion can give those regions disproportionate influence. Text-to-image attention offers an initial remedy by querying visual patches with linguistic states, but attention alone does not determine how strongly the returned context should enter each textual position. Our alignment stage therefore adds an adaptive gate: text states define the visual attention distribution, and a sigmoid control scales the resulting context before it is merged through a residual path.

The resulting model is a unified multitask formulation built around orthogonal representation decoupling. T5-base and CLIP ViT-B/16 encode language and vision; an adaptive module aligns their features; common and task-dependent projections divide the aligned state; and a prompt-conditioned decoder supports the three objectives. Training combines summary generation, image-text matching, image-text retrieval, and an orthogonal regularizer. At test time, however,

only the summarization behavior is activated. Matching and retrieval serve as sources of supervision during learning rather than as a separate retrieval procedure during generation.

This study makes three contributions. It first places summarization, matching, and retrieval in one model whose aligned multimodal encoder is optimized through a joint objective. It then introduces a shared-private decomposition with a Frobenius-norm penalty that discourages correlation between transferable and task-specific features, limiting negative transfer. Finally, it combines text-conditioned visual attention with gated denoising and evaluates the resulting design through baseline comparisons, component ablations, and a sensitivity analysis of λ4 on the datasets used in the experiments.

2 RELATED WORK

2.1 Multimodal Summary Generation

Initial work on multimodal summarization encoded an article and its associated image before producing a textual synopsis [1, 4]. Subsequent aspect-aware and hierarchical attention mechanisms improved how visual evidence was selected [5-6]. Unified vision-language architectures have more recently expressed several tasks as generation problems or adopted contrastive losses to narrow the semantic gap between modalities [7-9]. Collectively, these studies demonstrate the value of auxiliary cross-modal signals. Nonetheless, feature concatenation and hard parameter sharing remain common, leaving interference among task objectives largely unmanaged.

2.2 Learning Across Multiple Tasks

Multitask methods transfer information by sharing parameters, latent factors, or both across related objectives. Prompt-driven unified systems expose one parameter set to several task formulations [8, 10], whereas contrastive learning contributes pairwise signals that sharpen modal alignment. Other vision-language models likewise combine common generation mechanisms, pretrained alignment, and multimodal fusion to reuse supervision across tasks [11-18]. Here the auxiliary objectives are selected for distinct but complementary roles: matching contributes a binary judgment of semantic compatibility, while retrieval organizes the large-scale geometry of the joint embedding space.

2.3 Negative Transfer Across Objectives

Negative transfer describes the loss of performance that occurs when tasks require incompatible features or produce conflicting gradients. Hard sharing is economical, but it offers no structural protection for patterns unique to a particular objective. Factorized latent-variable models and domain separation networks instead motivate an explicit division between common and private information [2-3]. Following this line, we regularize the learned representations themselves: correlation between the shared feature matrix and each task-specific matrix is penalized directly.

2.4 Common and Task-Dependent Representations

In a shared-private architecture, a hidden state is separated into a component intended for transfer and additional components reserved for individual tasks. The former conveys information that can benefit several objectives; the latter prevent specialized variation from being propagated indiscriminately. Our framework instantiates one shared projection alongside three private projections for summarization, matching, and retrieval. A learned task token subsequently determines how these sources are recombined by the decoder.

2.5 Aligning Language with Visual Evidence

Cross-modal attention typically uses states from one modality as queries and representations from the other as keys and values. Although CLIP-style pretraining places text and images in a semantically comparable space [19-20], it does not make every visual patch relevant to a given article. The proposed alignment block goes beyond direct concatenation: it derives attention at the token level, estimates a joint text-vision gate, and introduces only the gated visual context through residual fusion. In this way, modal alignment is treated simultaneously as evidence selection and visual denoising.

3 PROBLEM FORMULATION

3.1 Multimodal Summarization

Let X={x1,…,xL} denote an article, I its associated image, and Y={y1,…,yT} the reference summary. The generation branch models P(Y|X,I). The article is represented by T5 as HT∈R^(L×768), while CLIP ViT-B/16 maps I to HV∈R^((N+1)×768), comprising N patch tokens and one global token. Conditioned on both streams, the decoder predicts Y autoregressively. Human-written summaries provide supervision, and performance is assessed with ROUGE-1, ROUGE-2, ROUGE-L, and BLEU [21-22].

3.2 Image-Text Compatibility

For matching, the input is an image-text pair (I,T) with a binary target l∈{0,1}. An observed pair is positive; an image combined with unrelated text is negative. The corresponding head, or decoder behavior, estimates pmatch, the probability that the two inputs are semantically consistent. Binary cross-entropy then rewards pair-level agreement in the shared representation and penalizes unsupported cross-modal associations.

3.3 Cross-Modal Retrieval

The retrieval task uses one modality to query candidates in the other. Within a batch of N aligned pairs, diagonal entries are treated as positives and all off-diagonal combinations as negatives. Cosine similarity is calculated between projected text and image vectors, after which an InfoNCE loss with temperature τ is applied. Unlike the token-level generation objective, this signal explicitly encourages a globally discriminative shared embedding.

3.4 Joint Optimization

A common set of multimodal encoders, the adaptive alignment block, and the shared-private projections supports all three tasks. Task-specific behavior is selected by Etask and realized through the appropriate prediction head or decoder mode. With Lgen, Lmatch, Lretr, and Lorth representing the generation, matching, retrieval, and orthogonality terms, respectively, learning minimizes a weighted combination of the four losses. Section 4.9 gives the coefficients used in the reported experiments.

4 PROPOSED METHOD

4.1 Architecture

The complete pipeline is depicted in Figure 1. Separate pretrained encoders first process the article and the image. Their outputs are passed to the adaptive cross-modal alignment block, which returns a denoised multimodal sequence. This sequence is then transformed into one common subspace and three task-specific subspaces. Guided by a learned task token, a unified decoder produces a textual summary, a matching prediction, or a retrieval representation. MSMO, Flickr30K, and MSCOCO supply the training data for the respective objectives. Once training is complete, inference retains only the summarization mode.

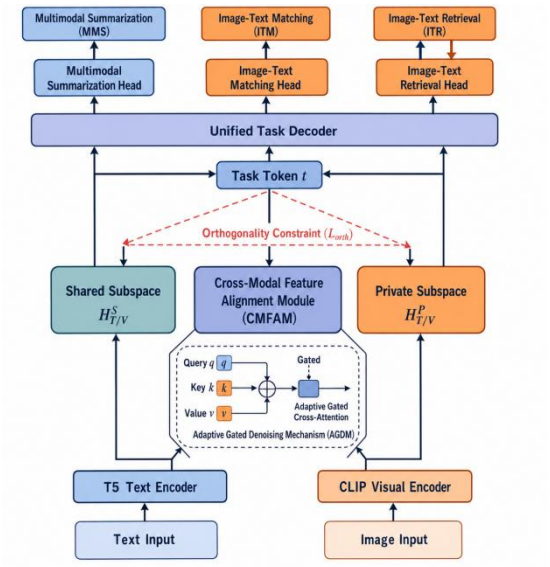

Figure 1 Multitask Architecture for Summary Generation, Image-Text Matching, and Cross-Modal Retrieval

4.2 Language and Image Encoding

Article tokens receive lexical and positional embeddings before passing through the 12-layer T5-base encoder. Its contextual output is HT∈R^(L×dtext), with dtext=768. The visual pathway partitions each image into non-overlapping 16×16 patches, prepends a trainable class token, and adds positional information before CLIP ViT-B/16 processing. It returns N local patch states together with a global state, all projected to dvis=768. Because dtext and dvis are equal, cross-modal attention can operate on the two sequences without an additional dimensional bridge.

$$\text{Etext} = \text{Embed(X)} + \text{PosEmbed(X)} \quad (1)$$

$$\text{HT} = \text{Encodertext(Etext)} \in \text{R}^{\wedge}(\text{L} \times 768) \quad (2)$$

$$\text{N} = (\text{H} \times \text{W})/(16 \times 16) \quad (3)$$

$$HV = Encodervis([xclass; x1E; ...; xNE] + Epos) \quad (4)$$

4.3 Text-Guided Adaptive Alignment

In the alignment block, language states serve as queries and visual states supply keys and values. The direction reflects the intended use: the evolving textual context should decide which regions of the image can substantiate the summary. As detailed in Figure 2, learned linear maps produce Q, K, and V. Scaled dot-product attention then assigns a distribution over visual tokens, from which the context CV is formed. A sigmoid gate λ is computed from the concatenation of HT and CV. The term λ·CV regulates visual contribution element by element, and a residual connection followed by layer normalization yields Haligned.

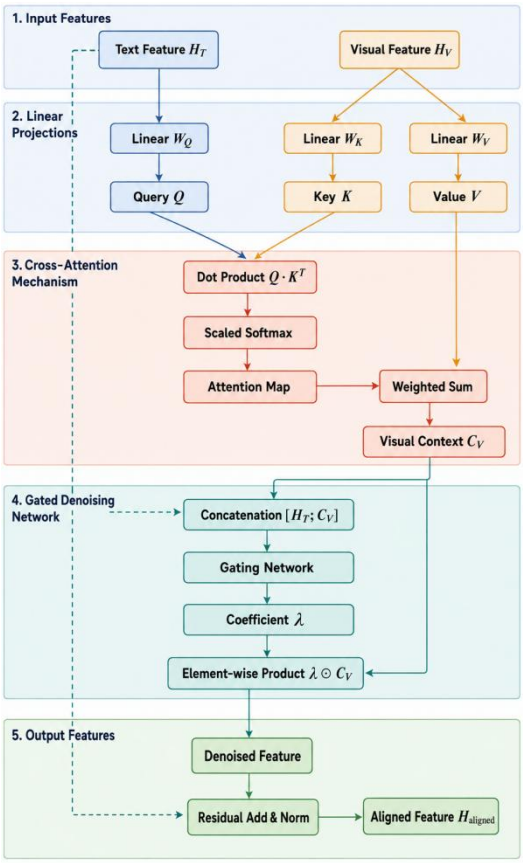

Figure 2 Text-Conditioned Visual Attention Followed by Adaptive Gating and Residual Denoising

$$Q = HTWQ, K = HVWK, V = HVWV \quad (5)$$
$$CV = softmax(QKT/\sqrt{dk})V \quad (6)$$
$$\lambda = \sigma(Wg[HT; CV] + bg) \quad (7)$$
$$Haligned = Linear(Concat(HT, \lambda \odot CV)) \quad (8)$$

Values of λ near one indicate that the attended image content adds useful evidence and should pass largely unchanged. When λ is close to zero, visual responses that repeat or contradict the current linguistic state are attenuated. Unlike fixed sequence concatenation, this operation adjusts the amount of injected visual information for each token according to its content.

4.4 Shared and Private Feature Spaces

Haligned is projected to a common representation HS and to three private representations Hp(m), where m∈{Sum, Match, Retr}. HS is intended to retain cross-modal meaning that transfers among tasks. The private branches instead preserve the properties that each objective needs: lexical and discourse structure for summarization, pair-level evidence for matching, and fine similarity distinctions for retrieval (Figure 3).

$$HS = GELU(HalignedWS + bS) \quad (9)$$
$$Hp(m) = GELU(HalignedWp(m) + bp(m)), m \in Sum, Match, Retr \quad (10)$$

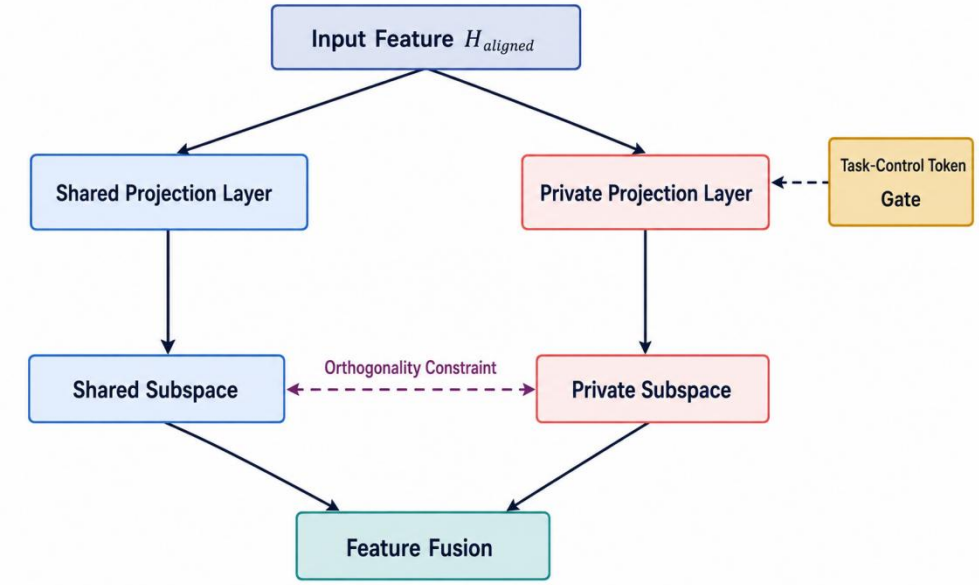

Figure 3 Separation of Common and Task-Specific Subspaces with Task-Aware Fusion

4.5 Orthogonal Decoupling

Distinct projection matrices do not by themselves stop a private branch from reproducing information already encoded in HS. We therefore impose a penalty on the correlation between HS and every Hp(m). The regularizer is defined as the squared Frobenius norm of their matrix product. Driving this quantity downward encourages near-zero inner products, assigning complementary information to the shared and private directions rather than permitting redundant copies.

$$Lorth = \Sigma m \in Sum, Match, Retr||HS^{T}Hp(m)||F^{2} \quad (11)$$

The strength of the constraint is set by λ4. If this coefficient is too small, the factorization remains weakly enforced; if it is excessive, geometric separation can dominate and suppress beneficial transfer. For the validation configuration reported here, the sensitivity analysis in Figure 4 reaches its highest ROUGE-L value at λ4=0.05.

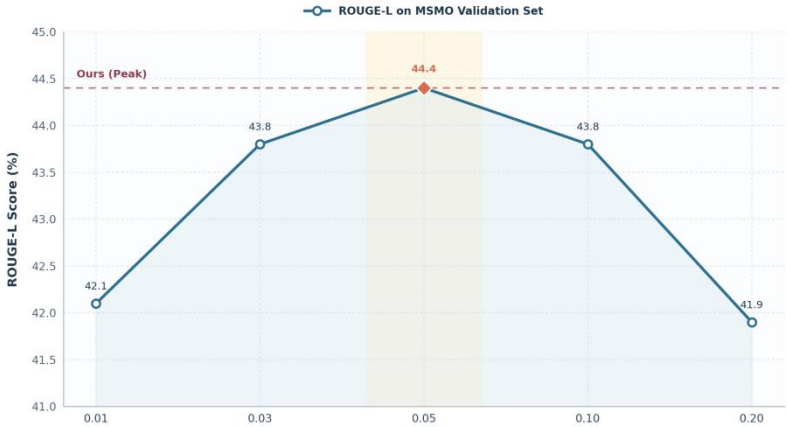

Figure 4 Validation ROUGE-L under Different Orthogonal-Loss Weights λ4

4.6 Summary Generation Branch

In summarization mode, the decoder receives HS, Hp(Sum), and the corresponding task token, attends over the aligned multimodal sequence, and emits the summary from left to right. Teacher forcing uses the human reference as the target sequence. Task-private features retain preferences concerning vocabulary and sentence organization, while the shared representation continues to provide aligned visual-language evidence.

$$Hfinal = HS + Hp(m) + Etask \quad (12)$$

$$Lgen = -\Sigma t log P\theta(yt|y < t, X, I; Etask = Sum) \quad (13)$$

4.7 Matching Branch

For image-text matching, the task token activates the matching-specific representation and prediction behavior. Given either a paired or mismatched input, the branch returns pmatch. The associated binary cross-entropy objective is defined as follows:

$$Lmatch = -E[llogpmatch + (1 - l)log(1 - pmatch)] \quad (14)$$

4.8 Retrieval Branch

For retrieval, projected image and text features occupy a common embedding space. The score for image i and text j is the cosine similarity sim(vI(i),vT(j)). Each diagonal image-text pair is positive, whereas the remaining within-batch pairs supply negatives. The image-to-text component of the InfoNCE objective is:

$$Lretr = -\Sigma ilogexp(sim(vI(i), vT(i))/\tau)/\Sigma jexp(sim(vI(i), vT(j))/\tau) \quad (15)$$

This objective contributes a ranking signal to the shared encoder. Although Equation (15) presents the image-to-text direction used in the implementation report, its underlying interpretation is bidirectional: aligned samples are drawn closer, while incompatible image-text combinations are separated.

4.9 Complete Training Loss

The four learning signals enter a single objective:

$$Ltotal = \lambda 1Lgen + \lambda 2Lmatch + \lambda 3Lretr + \lambda 4Lorth \quad (16)$$

Experiments use λ1=1.0, λ2=0.2, λ3=0.2, and λ4=0.05. This weighting preserves generation as the principal target while allowing matching, retrieval, and representation separation to influence the common encoder. Retrieval remains an auxiliary training objective and does not introduce an extra stage when summaries are generated.

4.10 Training and Inference

Training either alternates among or jointly samples task examples from MSMO, Flickr30K, and MSCOCO. Gradients from every available objective update the common encoders and the alignment module; each private projection and task head is updated by its own branch. The active operation is specified through the task token in the unified decoder. During evaluation, an MSMO article-image pair is encoded once and coupled with the summarization token. Neither the matching head nor the retrieval head participates in producing the output, although both have shaped the learned representation during training.

5 EXPERIMENTAL SETUP

5.1 Data

The unified experiment draws on three datasets. MSMO supplies news articles, associated images, and human-written summaries. Flickr30K, in which each everyday image has five human captions, supports the matching objective. MSCOCO provides the larger image-caption collection used for retrieval. Table 1 reproduces the train, validation, and test sizes used for all three branches.

Table 1 Dataset Splits for the Three Learning Objectives

Task	Dataset	Train	Validation	Test
Summarization	MSMO	293,965	10,355	10,261
Image-text matching	Flickr30K	29,000	4,589	5,273
Image-text retrieval	MSCOCO	113,287	6,530	7,321

5.2 Comparison Systems

The evaluation covers Seq2Seq; the pointer-generator PG-Net; text-only T5; the multimodal Uni-ED and ATG-Transformer systems; the multitask or pretrained models UNIMO and UniSumm; and the larger vision-language model LLaVA-1.5-7B. This collection spans conventional single-task generation, multimodal fusion, unified multitask learning, and transfer from a large multimodal model.

5.3 Metrics

ROUGE-1 and ROUGE-2 quantify unigram and bigram overlap, respectively, whereas ROUGE-L relies on the longest common subsequence. BLEU combines clipped n-gram precision with a brevity penalty [21-22]. Every MSMO system is compared using these same automatic measures; no additional scores are inferred from the visualized results.

5.4 Implementation

All experiments run under Linux using PyTorch and HuggingFace Transformers on an NVIDIA GeForce RTX 3090. The T5-base text encoder contains 12 Transformer layers, uses a hidden dimension of 768, and has 12 attention heads. CLIP ViT-B/16 forms the visual encoder, operating on 16×16 patches and returning 768-dimensional projected features. AdamW is used for optimization [23]. Configuration details absent from the underlying experiment are deliberately left unspecified.

5.5 Loss Coefficients

The multitask coefficients are fixed at λ1=1.0, λ2=0.2, λ3=0.2, and λ4=0.05. The orthogonality analysis evaluates λ4 values of 0.01, 0.03, 0.05, 0.10, and 0.20. Other hyperparameters follow the reported configuration of the third-chapter experiment.

6 RESULTS AND DISCUSSION

6.1 Aggregate Results

Results on the MSMO test set appear in Table 2. The complete framework achieves 46.95 for ROUGE-1, 22.03 for ROUGE-2, 44.42 for ROUGE-L, and 18.71 for BLEU. Against Seq2Seq, the gains are 8.42 points in ROUGE-1 and 8.58 points in ROUGE-L; relative to text-only T5, the corresponding increases are 4.80 and 5.57 points. The ROUGE-L advantage over UNIMO and UniSumm is 3.19 and 2.57 points, while BLEU rises by 2.84 and 2.37 points. These outcomes show that simply providing an image is not the decisive factor. Stronger performance emerges when matching and retrieval supervision are combined with selective alignment and task decoupling, improving coverage without sacrificing fluency (Figure 5).

Table 2 Performance on the MSMO Test Set

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Seq2Seq	38.53	16.21	35.84	12.15
PG-Net	40.25	17.53	37.12	13.48
T5	42.15	18.20	38.85	14.35
Uni-ED	42.71	18.94	39.45	14.62
ATG-Transformer	43.15	19.42	40.08	15.13
UNIMO	44.52	20.15	41.23	15.87
UniSumm	45.18	20.76	41.85	16.34
LLaVA-1.5-7B	45.80	21.15	43.08	17.10
Ours	46.95	22.03	44.42	18.71

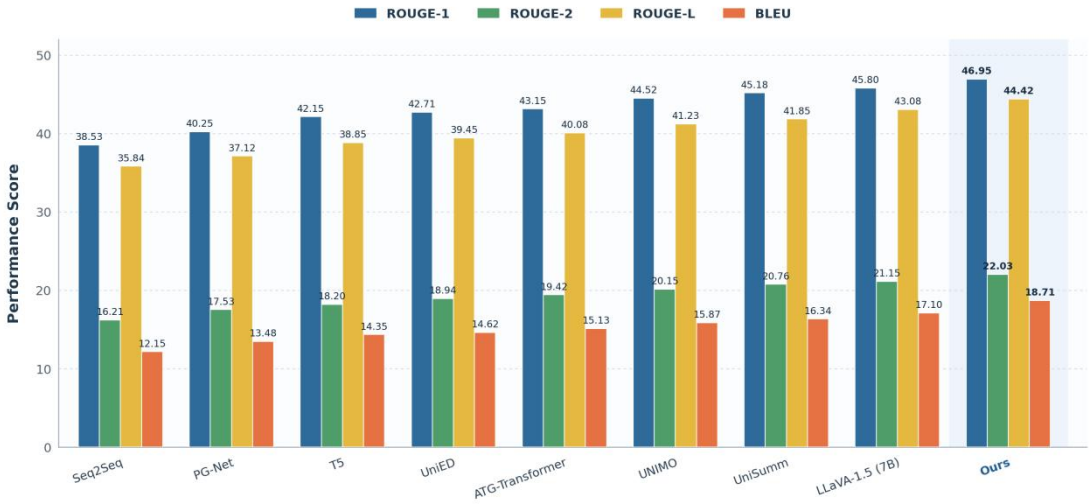

Figure 5 ROUGE and BLEU Scores of the Evaluated Baseline and Multitask Systems

6.2 Contribution of Auxiliary Tasks

Matching and retrieval shape the encoder in different ways. The matching loss teaches it to distinguish semantically coherent image-text pairs from incompatible ones, reinforcing grounding at the level of a complete pair. Retrieval adds a batch-wise ordering constraint and regularizes the broader geometry of the shared space. Consequently, the summarizer inherits a representation informed by both local compatibility and global discrimination. Improvements over UNIMO and UniSumm further suggest that shared parameters alone are insufficient; the private projections and their orthogonal regularization help convert auxiliary supervision into useful transfer rather than interference.

6.3 Component Ablations

Table 3 separates the effects of the three central design choices. Without adaptive alignment, ROUGE-L falls from 44.42 to 41.51, a decline of 2.91 points, and BLEU decreases from 18.71 to 16.85. Removing the shared-private factorization produces 42.34 ROUGE-L and 17.62 BLEU, including a 2.08-point loss in ROUGE-L. When only the orthogonal penalty is omitted, the scores become 43.55 and 18.24; the 0.87-point ROUGE-L reduction is smaller but remains consistent. The pattern is revealing. Direct filtering of irrelevant visual content contributes most, while subspace separation and geometric regularization provide additional safeguards against interference among tasks.

Table 3 Ablation Results on MSMO

Model	ROUGE-1	ROUGE-2	ROUGE-L	Δ(R-L)	BLEU
Full Model	46.95	22.03	44.42	–	18.71
w/o Alignment	43.82	19.56	41.51	−2.91	16.85
w/o Shared-Private	44.75	20.41	42.34	−2.08	17.62
w/o Orthogonal Loss	45.88	21.25	43.55	−0.87	18.24

6.4 Role of Orthogonal Separation

The variant without the orthogonal loss retains the common-private projections but fixes λ4 at zero. Its 0.87-point decrease in ROUGE-L indicates that separate projection layers do not remove all representational overlap. The Frobenius penalty supplies a direct learning signal against private features that replicate shared directions. This observation supports a representation-level account of the improvement, although it should not be interpreted as proof that semantic independence is complete.

6.5 Role of Adaptive Cross-Modal Alignment

Eliminating the alignment block causes the largest deterioration in the ablation study. Under direct concatenation, every image token is exposed to the decoder, including patches unrelated to the article. The full model instead uses language states to locate pertinent regions and a learned gate to regulate the amount of attended context admitted to the fused sequence. Cleaner inputs then reach both the shared and private projections, a mechanism consistent with the observed gains in ROUGE and BLEU.

6.6 Sensitivity to λ4

Figure 5 changes λ4 while holding the other settings constant. Validation ROUGE-L rises from 42.1 at 0.01 to 43.6 at 0.03, peaking at 44.4 when λ4=0.05. Larger coefficients reverse the improvement: scores decline to 43.8 at 0.10 and 41.9 at 0.20. A moderate orthogonality pressure is therefore preferable. Too little leaves common and private features entangled, whereas too much limits information that the tasks can profitably share.

6.7 Limitations

Several constraints remain. The relative influence of the four losses is governed by λ1–λ4 and may need to be recalibrated when task distributions change. The common representation is also sensitive to auxiliary-data quality; incorrect matches or uninformative retrieval negatives can produce misleading gradients. Moreover, the orthogonal term regulates linear correlation rather than guaranteeing semantic independence. Alignment quality depends on the language signal as well, so incomplete or ambiguous text may guide attention to unsuitable regions. Finally, learning three objectives increases optimization complexity, and generalization beyond the reported combination of MSMO, Flickr30K, and MSCOCO has yet to be established.

7 CONCLUSION

This paper has introduced a unified model that learns multimodal summarization together with image-text matching and retrieval. Its architecture joins T5 and CLIP encoders with text-guided visual alignment, common-private feature projections, and a Frobenius-norm orthogonality constraint. Generation, matching, retrieval, and subspace decoupling are weighted by λ1=1.0, λ2=0.2, λ3=0.2, and λ4=0.05. The full system obtains ROUGE-1/2/L scores of 46.95/22.03/44.42 and a BLEU score of 18.71 on MSMO. Ablations confirm separate contributions from adaptive alignment, the shared-private factorization, and orthogonal regularization. The central finding is that multimodal generation benefits when irrelevant visual evidence is filtered before decoding and transferable semantics are kept distinct from the features demanded by individual tasks.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zhu J, Li H, Liu T, et al. MSMO: Multimodal summarization with multimodal output. In: Conference on Empirical Methods in Natural Language Processing, 2018: 4154-4164.
[2] Salzmann M, Ek C H, Urtasun R, et al. Factorized orthogonal latent spaces. In: International Conference on Artificial Intelligence and Statistics. PMLR, 2010: 701-708.
[3] Bousmalis K, Trigeorgis G, Silberman N, et al. Domain separation networks. In: Advances in Neural Information Processing Systems, 2016.
[4] Hu J, Kundu S, Poria S, et al. UniMS: A unified framework for multimodal summarization with knowledge distillation. In: AAAI Conference on Artificial Intelligence, 2022.

[5] Li H, Zhu J, Zhang C, et al. Aspect-aware multimodal summarization for Chinese e-commerce products. In: AAAI Conference on Artificial Intelligence, 2020: 8188-8195.
[6] Zhang L, Wei H, Cheng J, et al. Hierarchical cross-modal attention for multimodal summarization. IEEE Transactions on Multimedia, 2023, 25: 1230-1241.
[7] Wang Z, Yu J, Yu A W K, et al. SimVLM: Simple visual language model pretraining with weak supervision. In: International Conference on Learning Representations, 2022.
[8] Chung J Y, Lei J, Tan H, et al. Unifying vision-and-language tasks via text generation. In: International Conference on Machine Learning. PMLR, 2021: 1931-1942.
[9] Liu N, Joty S. UniSumm: Unified contrastive learning for multimodal summarization. In: Annual Meeting of the Association for Computational Linguistics, 2023.
[10] Liu Y, Zhu W. Unified multi-modal learning for cross-task generalization. In: IEEE Conference on Computer Vision and Pattern Recognition, 2023: 1821-1830.
[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems, 2017: 5998-6008.
[12] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 2020, 21(1): 5485-5551.
[13] Chen J, He X, Guo Y, et al. CLIP-It! Language-guided video summarization. In: Advances in Neural Information Processing Systems, 2021.
[14] Lu J, Batra D, Parikh D, et al. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems, 2019.
[15] Li J, Li D, Xiong C, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, 2022.
[16] Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based image description evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4566-4575.
[17] Wei Y, Zhang L, Wang Y. Cross-modal information fusion for multimedia summarization: A survey. Journal of Computer Science and Technology, 2023, 38(1): 1-20.
[18] Jégou Y, Johnson J, Douze H. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 2019, 7(3): 535-547.
[19] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, 2021: 8748-8763.
[20] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations, 2021.
[21] Lin C Y. ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, 2004: 74-81.
[22] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation. In: Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.
[23] Loshchilov I, Hutter F. Decoupled weight decay regularization. In: International Conference on Learning Representations, 2019.