Online ISSN: 3104-9338
Print ISSN: 3104-932X

Entity-Guided Contrastive Learning for Multimodal Abstractive Sentence Summarization

Bin Zhou[1*], Zhiming Chen[1], Guihua Gao[1], Hao Zeng[1], Sai Du[1], Hanyan Cai[1]

[1] Yiyang Normal College, Yiyang, Hunan, China

Received: Oct 23, 2026

Revised:Oct 29, 2026

Accepted: Oct 30, 2026

Published online: November,5 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 12 (December 2026)

[*] Corresponding Author:Bin Zhou (1649840723@qq.com)

Abstract. Multimodal abstractive sentence summarization condenses a news sentence into a headline with the help of its accompanying photograph. Vision-guided models inject a global visual representation into the encoder, yet they treat the entities named in the source sentence as ordinary tokens, so the decoder receives little fine-grained grounding and the visual signal is easily swamped by the stronger linguistic signal. This paper presents an entity-guided contrastive framework built on BART. Knowledge-graph entities recognised in the source sentence are represented with pretrained TransE embeddings and encoded jointly with the text and with the image by two encoders that share parameters. A gated fusion module combines the two encoder outputs into a single memory for the decoder, and contrastive objectives align the resulting representations with each other and with the target-summary representation. On the 2,000 instances of the MMSS test set, and against a text-only variant trained from the same pretrained weights under the same recipe, the framework improves ROUGE-1 from 55.13 to 56.38, ROUGE-2 from 31.60 to 32.89 and ROUGE-L from 51.38 to 52.47, and a paired bootstrap places the ROUGE-1 gain outside the zero-difference interval. A diagnostic section locates the remaining headroom in the decoder rather than in the alignment: the contrastive representations do separate the two branches, but the decoder grants the visual block under two per cent of its cross-attention mass, which bounds what the alignment can contribute to the headline.

Keywords: *multimodal summarization; abstractive sentence summarization; entity-guided learning; contrastive learning; cross-modal alignment*

1. Introduction

Online news is published as a single sentence with a photograph, and readers expect a headline capturing the central event of both. Multimodal abstractive sentence summarization (MAS) formalises this: given a source sentence and its photograph, the system produces a shorter, abstractive, visually consistent headline [1,2]. Since a headline is often all a reader sees, small content-selection differences change how much of the event survives.

Text summarization advanced through three generations of models. Early abstractive systems encoded the source with a convolutional network and decoded with an attention-based language model [3]; pointer-generator networks added a copy mechanism for rare words [4]; and pretrained encoder-decoders such as BART [5], PEGASUS [6] and ProphetNet [7] replaced task-specific architectures with large-scale denoising pretraining. On short-input benchmarks these models already produce fluent headlines, and remaining errors are predominantly of content selection rather than fluency.

Visual evidence can correct such errors, which is why MAS attracts attention. The most widely used backbone is VG-BART [8], inserting cross-modal attention into the BART encoder so textual states attend to visual patches. Related designs include the coarse-to-fine network CFSum [9], multi-stage fusion with a forget gate [10], and summary-oriented vision modelling [11]. Earlier work extended the task to joint summary and image generation in MSMO [12] and to video in How2 [13]. These models share one choice: the image enters the encoder as a global, unsegmented visual context.

Two limitations follow. First, a news photograph contains several objects, only some named in the sentence, and a global visual representation cannot indicate which to preserve. Second, the entities in the source are encoded like ordinary words, so no dedicated route lets entity-level knowledge reach the decoder. Entity-guided multimodal summarization (EGMS) [14] addressed the second by retrieving knowledge-graph (KG) entities and encoding them with text and image. Its gains confirm entity information helps, but the three modalities are still concatenated into one sequence and processed by a shared encoder whose attention is distributed by vector magnitude as much as by relevance.

Contrastive learning controls that distribution. CLIP [15] and ALBEF [16] align image and text by maximising similarity of matched pairs against in-batch negatives; BLIP [17] adds a momentum encoder and a queue of negatives; SimCLR [18] established that augmentation composition and negative-set size determine behaviour; and Align-and-Attend [19] applies contrastive objectives to video summarization at the sample level. These are instance-level:

they ask whether an image and text belong together, not whether a particular entity is grounded in the visual evidence.

This paper therefore studies an entity-guided contrastive formulation of MAS, in which entity, textual and visual representations are aligned at the representation level before the decoder consumes them. The contributions are as follows.

(1) An entity-guided encoding scheme represents KG entities with pretrained TransE embeddings [20] and encodes them jointly with text and image by two parameter-sharing encoders, so entity information is available on both the linguistic and visual side.

(2) Contrastive objectives are adapted: an entity-text against entity-image loss aligning the two encoder outputs, and an instance-level objective anchoring both to the target-summary representation.

(3) The framework is evaluated on MMSS [1] against a text-only variant sharing its backbone, data split, optimiser and labelling, so the reported difference is attributable to entity and visual memory rather than the training recipe, and a diagnostic reports how much contrastive alignment survives into the headline.

2. Related Work

2.1. Text Summarization

Methods differ in how aggressively they compress. Extractive approaches rank source units by centrality, as in LexRank [21], or learn to score sentences, as in SummaRuNNer [22]; sentence compression deletes words under syntactic constraints [23]; and selective encoding [24] learns which source tokens to keep. These guarantee grammaticality by reusing source spans but cannot paraphrase, which headlines often require.

The adoption of pretrained encoder-decoders changed the dominant paradigm. BART [5] showed a denoising objective yields a strong summarization prior; PEGASUS [6] specialised it to gap-sentence generation; and ProphetNet [7] introduced n-stream self-attention for multi-token prediction. On sentence-level benchmarks the gap between best systems is now small, and remaining errors concern which facts survive compression rather than fluency. This motivates the present work, which adds an external source of factual salience rather than a stronger decoder.

2.2. Multimodal Summarization

UzZaman et al. [25] presented an early formulation generating the summary of a complex

sentence with reference to a temporal image sequence, establishing that the image is consulted at encoding time. With MMSS [1], which pairs each sentence with its original photograph, sentence-level MAS became a standard benchmark, followed by a series of fusion architectures. MAtt applies modality attention to reweight visual features [1]; Doubly-Attentive interleaves text and image attention; MSE introduces a modality-specific encoder; UniG treats the image as a generation guide; VG-BART inserts cross-modal attention into the pretrained encoder [8]; and CFSum models the coarse-to-fine contribution of each modality and token [9]. All are compared under one protocol in Section 4.

A second line of work changes the output or input modality. MSMO [12] generates an image with the summary, learning cross-modal selection in both directions; How2 [13] summarises instructional videos with the visual stream temporally aligned to speech; and summary-oriented vision modelling [11] trains the visual encoder from a summary-derived signal. The setting here is the simplest: one sentence, one image, one headline. The restriction is deliberate, isolating how entity-level information should be injected.

A common shortcoming of the above fusion architectures is that the visual signal is injected as a whole. Whether a summary mentions a stadium, a player or a scoreline is decided implicitly by attention weights over patches, and the training objective never tells the model which image parts correspond to the entities a headline should preserve. Entity-guided summarization addresses this gap.

2.3. Entity-Guided Summarization

Entities carry most information a headline must retain, and large knowledge graphs such as ConceptNet [26] provide typed, disambiguated representations. EGMS [14] was the first to exploit this for MAS: it extracts entities by string matching against a filtered ConceptNet subgraph, represents them with TransE embeddings [20], and encodes an entity-text and an entity-image sequence with two parameter-sharing encoders. A sentence-entity gate then produces a convex combination on the shared entity slots, and the concatenation of the combined entity, text and image blocks is passed to the decoder as memory.

Two properties of that design matter here. First, the entity, text and image representations are produced by independently initialised projection matrices, so their magnitudes need not agree; once they share one encoder, attention is partly determined by magnitude rather than relevance. Second, the only supervised signal on the encoder outputs is the indirect generation loss: the decoder can produce acceptable headlines attending to the text block alone, leaving the entity and image blocks little gradient. Contrastive objectives provide a direct signal on exactly those

representations, motivating the framework studied here.

2.4. Contrastive Learning for Cross-Modal Alignment

Contrastive objectives learn representations by pulling matched pairs together and pushing mismatched pairs apart. SimCLR [18] showed that augmentation composition, temperature and negative-set size determine objective quality; CLIP [15] transferred the recipe to web-scale image-text pairs, yielding zero-shot transferable representations; ALBEF [16] aligned the unimodal encoders before fusing, so fusion operates on already comparable representations; and BLIP [17] stabilised the objective with a momentum encoder and a negative queue.

For summarization, Align-and-Attend [19] applies contrastive objectives to video summarization with multimodal output, reporting state-of-the-art results on four benchmarks (two video, two news). Its alignment is instance-level: a matched image-text pair is contrasted against other batch pairs. Entity-level alignment is stricter, since a sentence and image can belong together while the specific entity the headline mentions is absent from the picture. The two objectives are complementary, and the framework combines them.

3. Method

3.1. Problem Definition

Let $T = (t_1, ..., t_m)$ be a source sentence of m tokens, p the published photograph and $K = (k_1, ..., k_{L_e-1})$ the entity identifiers from matching T against a filtered ConceptNet subgraph. The task is to produce a headline $S = (s_1, ..., s_n)$, with $n < m$, preserving the central event of T and consistent with p. The model maximises the conditional likelihood of the reference headline given the sentence and photograph,

$$P(S \mid T, p) = \prod_{j=1}^{n} p(s_j \mid s_{<j}, T, p) \quad (1)$$

At inference the headline of maximum length-normalised likelihood is selected by beam search. The identifiers index pretrained TransE vectors [20], so co-occurring entities start from neighbouring points in the embedding space, allowing entity information to generalise beyond the training sentences in which an entity appears.

The notation follows EGMS [14]. An encoder-state block carries its modality as a superscript—E for entity slots, T for text, V for the visual block, D for the decoder state read by the matching head—and its branch as a subscript, e_t for the entity-text pass and e_i for the entity-image pass. The text block at the entity-text encoder output is therefore h_{et}^{T}, and the full output is $[h_{et}^{E}, h_{et}^{T}]$. Embeddings carry the modality alone, as e_t, e_e, e_v; encoders are f, projection

matrices W, and the gate weight w. The photograph is p with no subscript, whereas the distributions of Section 3.3 always carry one, as p_{ij}, p_i. Figure 1 uses these letters throughout, so every equation quantity can be located in the figure under the same name.

3.2. Model Architecture

The architecture is shown in Figure 1. An instance consists of a source sentence T, its photograph p and the entity set K extracted from T; the reference is S. All three inputs become vector blocks of width d: an entity block of L_e slots, a text block of at most L_t tokens and a visual block of L_v tokens. One BART encoder reads the entity block concatenated with the text block, then the same entity block with the visual block; a gated fusion combines the two outputs into a memory of $L_e + L_t + L_v$ tokens that the decoder attends to. The blocks, passes and gate follow EGMS [14]; the calibration below and the contrastive terms of Section 3.3 are this paper's contribution. Lengths and widths are symbolic and take values in Section 4.2.

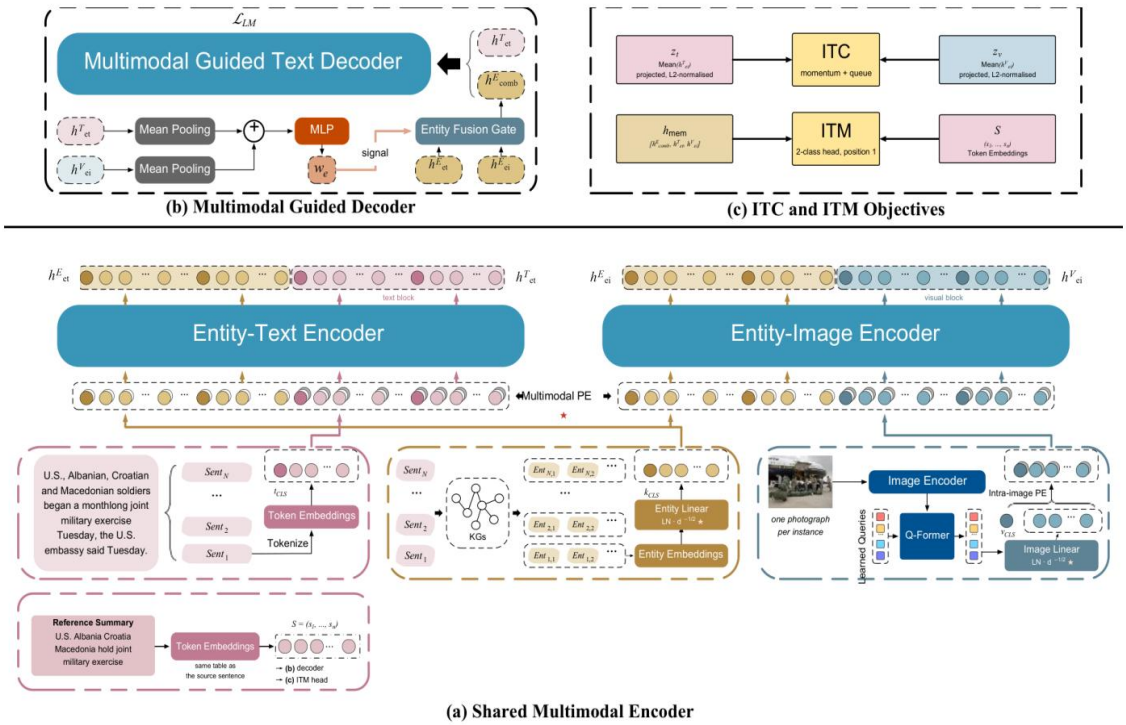


Figure 1. Architecture of the proposed framework, in the layout of EGMS [14]. (a) The shared multimodal encoder on one MMSS instance: T is embedded by W_t of Eq. (2), entities are looked up in W_{e1} and lifted by W_{e2} of Eq. (3), and p is read by $f_{img-enc}$, compressed by $f_{QFormer}$ and lifted by W_v of Eqs. (4) and (5); the encoder is applied twice with shared weights, giving h_{et} and h_{ei} of Eq. (8); the lower lane is the reference S that the decoder produces and the ITM head scores. (b) The gate of Eqs. (9) to (11): h_{et}^T and h_{ei}^V are pooled into w_e, combining h_{et}^E with h_{ei}^E; the brace marks the two blocks read beside the visual block of (a), and (b) writes h_{mem} of Eq. (12). The starred calibration of Eq. (6) is the modification introduced here. (c) The two contrastive objectives: ITC projects the two blocks to z_t and z_v of Eq. (14) and contrasts them against a momentum copy and a queue, and ITM classifies h_{mem}

against *S*.

The source sentence is embedded by the BART token table, with the delimiter tokens that EGMS uses to mark the two ends of the sequence,

$$e_t = W_t\,[\,t_{CLS}, t_1, \ldots, t_m, t_{SEP}\,] \tag{2}$$

Here W_t is the embedding matrix and the bracket denotes concatenation along the sequence dimension. Entities are obtained by matching *T* against a filtered ConceptNet subgraph [26], keeping at most $L_e - 1$ identifiers in order of first appearance. A lookup table initialised from TransE [20] maps an identifier to a vector, and a trainable map lifts it to the encoder width.

$$e_e = W_{e2}\,W_{e1}\,[\,k_{CLS}, k_1, \ldots, k_{L_e-1}\,] \tag{3}$$

Here W_{e1} is the frozen lookup table, W_{e2} the trainable projection and k_{CLS} a scene token playing EGMS's role: it carries a sentence-level state rather than any particular entity. Freezing the table preserves the relational geometry TransE learned over the whole graph while the model decides which directions matter for a headline, and keeps the entity pathway independent of the training-split size, which is too small to fit a table of entity vectors from scratch.

The photograph is encoded by a ViT-B/16 backbone as in BLIP [17], returning one feature per patch. The patches with the largest class-token attention weight are retained with the class token, and a query transformer with learnable queries compresses them,

$$r = f_{img-enc}(p) \quad v = f_{QFormer}(q_1, \ldots, q_{Lv-1};\, r) \tag{4}$$

$$e_v = [\,v_{CLS}, W_v\,v\,] + e_{intra-pos} \tag{5}$$

Here $f_{\mathrm{img\text{-}enc}}$ is the ViT, *r* the retained patch features, f_{QFormer} the query transformer applied to them, W_v the following projection and $e_{\mathrm{intra\text{-}pos}}$ the positional term of the visual block. A single learnt prefix token v_{CLS} is prepended to the query outputs, fixing the visual block length at L_v regardless of aspect ratio. Within each layer the queries attend to the retained patches, then to each other, then pass through a feed-forward block, each step with a residual connection and layer normalisation. The ViT is fine-tuned from ImageNet and the query transformer trained from scratch; the entity table and the two BLIP contrastive projection heads are frozen.

One inherited property proved decisive: the relative magnitude of the three blocks. The projections are initialised independently, so nothing forces comparable scales. The entity and visual blocks therefore entered the encoder at roughly seven and fifteen times the norm of a text token, and the positional table, drawn from a normal distribution, added a further offset more than forty times as large and identical for every block token. Because Transformer attention logits and values scale with their input, the text block carried only a small, predictable fraction

of the sequence signal energy: the decoder could still produce fluent headlines from the text block alone while the entity and visual blocks dominated the residual stream. Each block is therefore calibrated before entering the encoder,

$$e_e \leftarrow \mathrm{LN}(e_e) \cdot d^{-1/2} \quad e_v \leftarrow \mathrm{LN}(e_v) \cdot d^{-1/2} \quad e_{multi-pos} \leftarrow 0 \tag{6}$$

where LN is layer normalisation over the model dimension, and the third expression is the multimodal positional table $e_{\text{multi-pos}}$ that Eq. (7) adds to both branches, with rows initialised to zero rather than a normal draw. Zero initialisation means a block starts contributing no offset, and the model decides from the generation loss how much it wants. The text block is deliberately left uncalibrated to stay in the pretrained encoder's expected range; after calibration the three blocks enter the encoder at one common scale.

The two branches then differ only in their suffix. Both start from the calibrated entity block, both add the multimodal positional table position by position, and both are read by the same BART encoder with shared weights,

$$e_{et} = [e_e, e_t] + e_{multi-pos} \quad e_{ei} = [e_e, e_v] + e_{multi-pos} \tag{7}$$

$$h_{et} = [h_{et}^E, h_{et}^T] = f_{enc}(e_{et}; \theta) \quad h_{ei} = [h_{ei}^E, h_{ei}^V] = f_{enc}(e_{ei}; \theta) \tag{8}$$

Here e_t and e_v are the blocks of Eq. (2) and (5), $e_{\text{multi-pos}}$ the table added block by block, and e_t, e_i the two input sequences of lengths $L_e + L_t$ and $L_e + L_v$, where $L_t = m + 2$ counts the two delimiter tokens of Eq. (2); h_{et} and h_{ei} are the encoder outputs, same lengths and width as inputs. The two passes share one parameter set θ, so outputs split identically: entity slots occupy the same positions on both sides, h_{et}^T follows them in the first branch and h_{ei}^V in the second.

Sharing the encoder keeps the parameter count at that of a single BART and gives the design a property Section 4.6 returns to: the two branches enter from the same entity embedding and positional row, so h_{et}^E and h_{ei}^E are two readings of one slot set rather than unrelated representations. They are not numerically equal, because the entity slots attend to a text suffix on one side and a visual suffix on the other, which makes a gate between them informative. The shared prefix also keeps the BART positional embeddings meaningful for both sequences.

The gate is Eqs. (4) to (6) of EGMS [14] applied to the differing blocks. It pools the first branch's text block and the second branch's visual block with a mask-normalised mean, concatenates the pooled vectors, and passes them through a two-layer perceptron with a rectified linear unit and sigmoid,

$$h_{tv} = \mathrm{Mean}(h_{et}^T) \oplus \mathrm{Mean}(h_{ei}^V) \tag{9}$$

$$w_e = \sigma(W_{2w}\,\mathrm{ReLU}(W_{1w}\,h_{tv} + b_{1w}) + b_{2w}) \tag{10}$$

$$h^E_{comb} = w_e\,h^E_{et} + (1 - w_e)\,h^E_{ei} \tag{11}$$

Here W_{1w}, W_{2w}, b_{1w} and b_{2w} are the parameters of a two-layer perceptron with a single output. The gate therefore yields a single scalar w_e per example rather than one per entity slot, and h^E_{comb} of Eq. (11) is the combined entity block, of length L_e like h^E_{et} and h^E_{ei}. A scalar gate is coarse but the simplest device expressing a branch preference, and it keeps gate parameters independent of the entity count. The memory concatenates the combined entity block with the first branch's text block and the second branch's visual block,

$$h_{mem} = [\,h^E_{comb},\, h^T_{et},\, h^V_{ei}\,] \tag{12}$$

giving $L_e + L_t + L_v$ tokens of width d, needing no further projection since the decoder has the encoder width. The decoder is a BART decoder sharing the token embedding and vocabulary with the encoder, so a source and a summary token are the same object in one space, and cross-attention over the memory is the only route by which entities and the image reach the output.

3.3. Training Objective

Training combines a generation loss with two contrastive terms. The generation loss is the usual token-level cross-entropy,

$$\mathcal{L}_{LM} = -\frac{1}{n}\sum_{j=1}^{n} \log p(s_j \mid s_{<j}, h_{mem};\, \varphi) \tag{13}$$

where s_j is the j-th reference token and $s_{<j}$ the preceding tokens, so each term is the negative log probability the decoder assigns to the correct token given those generated; the conditioning is on the memory h_{mem} of Section 3.2 rather than the source sentence, φ collects the model parameters, and the sum runs over the n target positions and is divided by their number. No label smoothing is applied.

The first contrastive term acts on the two encoder outputs before fusion. Each branch contributes only its own modality: the text block of the entity-text branch and the visual block of the entity-image branch are pooled with the same mask-normalised mean, while the common entity block is excluded. Including it would let the objective be satisfied by recognising that two examples share an entity, an easier problem than deciding whether a sentence and a photograph belong together. The two pooled vectors are then carried into a common space by two separate linear maps and L2-normalised,

$$u_t = \mathrm{Mean}(h^T_{et}) \quad u_v = \mathrm{Mean}(h^V_{ei}) \quad z_t = W_{ct}\,u_t \quad z_v = W_{cv}\,u_v \tag{14}$$

and the similarity of a text-side query to an image-side candidate is the cosine of the two projected vectors,

$$a_{ij} = \frac{z_{t,i} \cdot z_{v,j}}{\|z_{t,i}\| \, \|z_{v,j}\|} \tag{15}$$

which is turned into a distribution over the N candidates by

$$p_{ij} = \frac{\exp(a_{ij}/\tau)}{\sum_{k=1}^{N} \exp(a_{ik}/\tau)} \tag{16}$$

where τ is a learned temperature clamped to a fixed interval; B the batch size, Q the queue length and $N = B + Q$ the number of candidates, given in Section 4.2. All N are produced by a momentum copy of the encoder and both projections, updated by an exponential moving average and evaluated with dropout disabled: the Q queued representations were written by earlier steps and the B in-batch ones are recomputed through the same momentum copy, so queries and candidates share one space. The queue makes the objective estimable since a single batch gives only one negative per query, and its length is rounded up to a multiple of B so the ring buffer never straddles two batches. Following BLIP [17], the objective uses momentum distillation: the same momentum copy scores every candidate, giving a_{ij}^{mom}, and the same temperature softmax yields a teacher distribution p_{ij}^{mom} smoother than the student one because the momentum weights average over past steps. The target is thus not one-hot but a mixture of that teacher and the identity,

$$q_{ij} = (1 - \alpha)\, \delta_{ij} + \alpha\, p_{ij}^{mom} \tag{17}$$

where α is a mixing weight, δ the identity (one at $j = i$, zero elsewhere), and the teacher supplies the remaining mass. The soft part stops the objective treating two examples describing the same event as mutually exclusive, common on this repetitive corpus. Both directions are computed and averaged,

$$\mathcal{L}_{ITC} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\sum_{j=1}^{N} q_{ij}^{T} \log p_{ij}^{T} + \sum_{j=1}^{N} q_{ij}^{V} \log p_{ij}^{V} \right] \tag{18}$$

The second term is an instance-level matching objective over the fused memory. Positive pairs are a headline with its own example's memory. Negatives are built two ways - pairing a headline with a different example's memory and pairing an example's memory with a different headline - where the swapped example is drawn from a categorical distribution proportional to the softmax of the current ITC similarity, making the negatives the batch's hardest. The three sets pass through the decoder once with bidirectional attention, and the first-position hidden state, written h_i^D for pair i, is the joint representation. Following BLIP [17], the decision uses a two-way linear head trained with cross-entropy rather than thresholding similarity, so the head

learns its own boundary instead of inheriting one from the contrastive space:

$$p_i = \text{softmax}(W_c\, h_i^D + b_c) \quad \mathcal{L}_{ITM} = -\frac{1}{3B}\sum_{i=1}^{3B} \log p_{i,yi} \tag{19}$$

where h_i^D is the first-position hidden state, read after the decoder processes the memory with its causal mask disabled so every position sees the whole headline; W_c is the two-way projection weights and bias, making the head the smallest trainable module; p_i is the two-class distribution over pair i's labels and $p_{\text{i,yi}}$ the entry for the true one. The head is used only during training and discarded at inference, where the decoder generates without it.

The three terms are combined as

$$\mathcal{L} = \lambda_{LM}\mathcal{L}_{LM} + \lambda_{ITC}\mathcal{L}_{ITC} + \lambda_{ITM}\mathcal{L}_{ITM} \tag{20}$$

where the λ coefficients weight the three terms. The auxiliary weights are deliberately small: the generation loss is the only term defined on the task's actual quantity, and the contrastive gradients come from a single batch, so a large weight would trade headline quality for a noisy alignment objective. All three are logged separately, enabling the Section 4.6 diagnostic.

4. Experiments and Analysis

4.1. Dataset

The framework is evaluated on MMSS [1], a Daily Mail derived corpus in which each source sentence is paired with its original photograph and the article headline as reference summary. The corpus contains 66,000 sentence-image-headline triples, split into 62,000 training, 2,000 validation and 2,000 test instances; Table 1 summarises the length distribution. Both sources and references are short: in training the source averages 21.7 words and the reference 7.7, with minima 11 and 2 and maxima 63 and 25 respectively, and validation and test are comparable. This short input is characteristic of the task and is why content selection, not compression ratio, dominates the error analysis.

Table 1. Statistics of the MMSS dataset. Lengths are given in words as minimum / mean / maximum.

Split	Size	Source length, words (min/mean/max)	Reference length, words (min/mean/max)
Train	62,000	11 / 21.68 / 63	2 / 7.72 / 25
Validation	2,000	11 / 24.35 / 47	3 / 7.68 / 17
Test	2,000	11 / 22.97 / 51	3 / 7.67 / 24

4.2. Experimental Settings

The symbolic sizes of Section 3.2 take these values. The entity block holds L_e = 48 slots, the

text block is capped at L_t = 96 tokens and the visual block at L_v = 33 (32 query outputs plus the prefix token); the ViT retains 64 patches with the class token; the model width is d = 768; and the two ITC projections map into 256 dimensions. These are architectural settings, not corpus properties: m and n, the lengths of a source sentence and its headline, vary per instance and are bounded rather than fixed.

The calibration of Eq. (6) equalises the three blocks by construction: before it the mean per-token norm of the entity and visual blocks is 7.03 and 15.47 against 1.07 for a BART token, and the inherited positional table adds an offset of norm 47.4; after it the three norms are 1.07, 1.00 and 1.00. Counting the embeddings, shared encoder, query transformer, decoder and contrastive heads, the model has 688.2 million parameters, of which 357.1 million are trainable and the rest frozen pretrained weights.

The three losses are weighted with λ_{LM} = 1.0, λ_{ITC} = 0.1 and λ_{ITM} = 0.1. The temperature of Eq. (16) starts at 0.07 and is clamped to [0.001, 0.5], the candidate-copy momentum is 0.995, and the target of Eq. (17) mixes α = 0.5 of the momentum distribution with the identity. The queue holds 2,060 rather than a round 2,048 representations, since its length is rounded up to a multiple of the batch size 20; with B = 20 this gives N = 2,080 candidates per query.

All text modules are initialised from facebook/bart-base [5] and the visual encoder follows the Vision Transformer design of BLIP [17]. Entity vectors come from TransE embeddings [20] trained on the filtered ConceptNet [26] subgraph released with EGMS [14], kept frozen during fine-tuning so the small training set does not overwrite the knowledge-graph structure. The model is trained with AdamW at learning rate 2e-5, weight decay 0.02, mini-batch size 20 and gradient accumulation over eight steps, giving an effective batch size of 160. A linear warm up over the first 1,500 optimizer steps precedes cosine decay to zero; the number of epochs is set per configuration, and no label smoothing is applied. Maximum source and target lengths are 96 and 40 tokens and the minimum target length is five. Decoding uses beam search with beam width five, length penalty 1.0 and trigram blocking, and the checkpoint with best validation ROUGE-1 is used for testing.

Because every configuration is trained from the same pretrained weights under the same recipe and run until its validation score stops improving, all Section 4.5 comparisons are between models differing only in the components under study. Random seeds are fixed and each configuration is trained once on a single NVIDIA RTX 2080 Ti GPU with 22 GB of memory. All experiments run offline with locally cached pretrained weights, so tokenisation and initialisation are identical across runs.

4.3. Evaluation Metrics

Following the standard MMSS protocol, ROUGE-1, ROUGE-2 and ROUGE-L are reported as F1 scores multiplied by 100 [27]. ROUGE-1 and ROUGE-2 measure unigram and bigram overlap with the reference, and ROUGE-L the longest common subsequence, rewarding longer contiguous matches without requiring adjacency. Porter stemming is applied before matching so morphological variants count as overlapping. BLEU-4 [28] is reported as a precision-oriented complement over whitespace-delimited words with add-one smoothing, so unseen n-grams do not zero the score; BERTScore F1 [29] captures semantic similarity that n-gram overlap misses. All metrics use the full 2,000 test instances, and the proposed framework's ROUGE implementation is verified against the reference on a fixed set of generations before any comparison.

4.4. Baseline Models

The text-only baselines are Lead, which copies the first source words; Compress [23], deleting words under syntactic constraints; ABS [3], the convolutional attentional encoder-decoder; and SEASS [24], selecting source tokens at encoding time. The multimodal baselines are Multi-Source, treating the image caption as an extra source; Doubly-Attentive and MAtt, applying attention over text and image [1]; MSE, with modality-specific encoders; UniG (T) and UniG, treating the image as generation guidance; CFSum [9]; and VG-BART [8]. Following MMSS convention, dagger-marked values are quoted from the original publications, all using BART-large backbones.

Two systems are deliberately not reproduced. VG-BART [8] pairs a cross-modal attention sublayer with a copy mechanism, and porting it into the present codebase - whose memory is built from entity, text and image blocks - would change several factors at once; it is discussed in Section 2 as the strongest published vision-guided backbone rather than tabulated. EGMS [14] is also omitted: its released configuration couples a BART-large backbone with a differently built entity graph, so a like-for-like re-run would mostly measure the training recipe. It remains the starting point of this design and is discussed as such in Section 2.

4.5. Main Results

Table 2 reports the automatic evaluation results. Among published baselines, vision-guided models outperform text-only ones, and the gap between CFSum [9] and earlier fusion architectures confirms that injecting visual evidence into a pretrained encoder beats fusing it after encoding. The table splits into two blocks deliberately: the upper quotes original

publications using BART-large backbones, while the lower reports this work under a BART-base backbone with the recipe of Section 4.2, so it is not directly comparable. With no published system under that recipe, this paper makes no claim of beating the quoted figures; it claims a controlled comparison: a same-recipe text-only variant fills the lower block's first row as the reference against which modality contributions are judged, and any multimodal configuration failing to clear it gains nothing from image or entities.

Table 2. Automatic evaluation results on the MMSS test set. Dagger marks values quoted from original publications using BART-large backbones and reporting BLEU-4 on a different tokenisation from the lower block's word-level BLEU-4; the lower block reports this work under a BART-base backbone, so the blocks are not directly comparable.

Model	R-1	R-2	R-L	BLEU-4
Reported in the original publications (BART-large backbones)				
Lead†	33.64	13.40	31.84	—
Compress†	31.56	11.02	28.87	—
ABS†	35.95	18.21	31.89	—
SEASS†	44.86	23.03	41.92	—
Multi-Source†	39.67	19.11	38.03	—
Doubly-Attentive†	41.11	21.75	39.92	—
MAtt†	47.28	24.85	44.48	—
MSE†	45.63	23.68	42.97	—
UniG (T)†	45.90	24.08	42.98	47.09
UniG†	46.22	24.28	43.47	46.85
CFSum†	47.86	25.64	44.64	48.83
This work, same recipe throughout (BART-base backbone)				
Text-only variant (10 epochs)	55.13	31.60	51.38	21.50
BLIP+EGMS (ours) (30 epochs)	**56.38**	**32.89**	**52.47**	**22.75**

That reference is the text-only variant. It is trained from the same pretrained weights with the same optimiser, learning rate, batch size, gradient accumulation and decoding settings, and differs from the proposed framework in two respects only: the knowledge-graph entities and the image are removed from the memory, so the encoder receives the source sentence alone, and no cross-modal contrastive objective is applied to single-modality representations. Against this reference the proposed framework improves ROUGE-1 from 55.13 to 56.38, ROUGE-2 from 31.60 to 32.89 and ROUGE-L from 51.38 to 52.47, and word-level BLEU-4 from 21.50 to 22.75, gains of 1.25, 1.29, 1.09 and 1.25 points respectively. On the same test instances, BERTScore F1 moves from 70.53 to 71.24.

The two rows differ in epochs: ten for the text-only variant and thirty for the complete framework. The text-only validation ROUGE-1 reached 54.56 at the seventh epoch, whereas

the complete framework was still climbing at the tenth, peaked at 55.23 at the thirteenth, and fluctuated within one point for the remaining seventeen. Neither run was stopped while training still bought accuracy, though the smaller validation gap of 0.67 points should be noted when reading the test difference. A paired bootstrap over 10,000 resamples of the 2,000 test instances places the ROUGE-1 difference at a mean of 1.26 points with a 95% confidence interval of [0.73, 1.80], excluding zero; the proposed framework wins on 484 instances, ties on 1,124 and loses on 392. The gain is small but consistent across the test set rather than driven by a few favourable examples.

Three properties support reading the gains as evidence about the modalities rather than noise. First, the gain is larger on ROUGE-2 than ROUGE-1 and smallest on ROUGE-L, so the multimodal memory adds better ordered bigrams rather than extra content words. Second, BLEU-4 moves with ROUGE and the mean generated length is 6.87 words against 6.86 for the text-only variant, so the improvement is not from longer headlines; the length penalty and minimum target length are fixed across runs. Third, comparing the proposed framework's generated headlines with their own sources gives ROUGE-1 of 44.18 against 44.73 for the text-only variant: the entity-guided model copies slightly less from the input while matching the reference slightly better, the signature of improved abstraction rather than extraction.

4.6. Analysis and Discussion

The central question of Section 2.3 is whether the three modalities' representations remain comparable after a shared encoder. Three diagnostics are reported, measured on the checkpoint whose contrastive term is computed per word rather than per sentence - the configuration in which alignment is actually learned. Its automatic scores lie within six tenths of a point of the Table 2 row, so what follows describes the same system.

First, the pairwise cosine similarity between the two encoder outputs is measured on forty test instances. Before the projection heads the mean similarity between a text-side and image-side sentence vector is 0.380, standard deviation 0.144 across pairs; after the projections and L2 normalisation it falls to 0.148. Neither value indicates collapse: the branches keep distinct directions and the contrastive loss is not on near-identical inputs. The same representations retrieve the paired image-side vector from a pool of forty at 67.5% top-1 and 100% top-5, against 2.5% for chance. The alignment is therefore real rather than nominal, which makes the null result below worth reporting.

Second, the decoder's attention over the three memory blocks is recorded on eight test instances. The entity block receives 30.5% of the mass across its 48 positions, the text block

67.7% across its 96, and the image block 1.7% across its 33, with entropy 3.419 against 5.176 for a uniform distribution. Normalised by valid positions the text block is by far the most attended: its 39 effective tokens take 0.0174 each, against 0.0064 for an entity slot. The ordering is expected, but the image's share bears on the interpretation below.

Third, the gate weight w_e of Eq. (10), averaged over the steps of each epoch, is monitored through training. It ends the first epoch at 0.50, reaches 0.66 by the fifth, stays between 0.60 and 0.69 for the remaining twenty-five, and closes at 0.60. It saturates at neither extreme, so neither branch is discarded; the equilibrium sits on the text side, consistent with the attention profile above.

Together, the three measurements describe a system whose cross-modal alignment succeeds at the optimised level yet does not reach the output: the contrastive term moves ROUGE-1 by less than half a point over the same recipe without it. The attention profile explains why: the image block carries under two per cent of the decoder cross-attention mass, so a better visual-textual match has almost no channel to influence the headline. The bottleneck is not alignment quality but the decoder's willingness to consult the visual block, and on a corpus whose source already contains the needed entities and predicate, attending to the photograph is not rewarded by the objective.

An error analysis examined the entity content of generated headlines. Since most of the 48 entity slots are irrelevant, the analysis is restricted to input entities that appear in the reference headline: 6,451 entities over 1,949 test instances. The generated headline retains 4,940, an entity recall of 76.6%, so 23.4% of entities a human writer kept are dropped, and in 99 instances (5.1% of the subset) every such entity is missing. Missing entities are the dominant error type, and losses concentrate on the modifier rather than the head of a name - karlsruhe and bundesliga dropped while dortmund survives, and college basketball replaced by a paraphrase. Hallucinated entities and wrong-event summaries would need manual annotation and are not quantified.

A limitation is that evaluation is restricted to a single corpus of short sentences. MMSS is the standard sentence-level MAS benchmark, but its headlines come from a small set of recurring event types, so a model can reach competitive ROUGE scores with a strong language prior and little grounding. Extending evaluation to longer inputs, such as the video summaries of How2 [13] or document-level summaries of MSMO [12], would test whether entity-level alignment matters more when the source contains many entities and the image only a few.

5. Conclusion

This paper studied an entity-guided contrastive formulation of multimodal abstractive sentence summarization. Entities recognised in the source are represented with pretrained TransE embeddings and encoded jointly with text and image by two parameter-sharing encoders; a gated fusion module combines the two outputs into the decoder memory; and contrastive objectives align the representations with each other. On the 2,000 MMSS test instances, against a same-recipe text-only variant, the complete framework improves ROUGE-1, ROUGE-2 and ROUGE-L by 1.25, 1.29 and 1.09 points and word-level BLEU-4 by 1.25 points, and a paired bootstrap places the ROUGE-1 gain outside the zero-difference interval. The improvement is modest: Section 4.6 shows the contrastive objectives move ROUGE-1 by less than half a point, the decoder grants the image block under two per cent of its cross-attention mass.

The magnitude mismatch of Section 2.3 proved a practical obstacle, not merely theoretical. With modality projections left as they are, the entity and image blocks can enter the shared encoder an order of magnitude larger than text tokens, so the decoder reads a memory where linguistic evidence is numerically negligible; normalising each modality block and zero-initialising the modality embedding removes that imbalance. The calibration is, however, applied by hand, and learning the per-modality scale is the obvious next step. Two further directions are promising: replacing the string-matching entity extractor with a trained linker so indirectly mentioned entities can be used, and extending the contrastive objective from sentence to entity level, aligning each entity with its image region.

Data Availability Statement

The MMSS dataset used in this study is publicly available from the authors of [1]. The filtered ConceptNet subgraph and the pretrained TransE embeddings are released together with EGMS [14]. No new data were created in this study.

Code Availability

The source code will be made publicly available upon acceptance.

Acknowledgements

This work was supported by the Scientific Research Fund of Hunan Provincial Education Department (General Program, Project No. 24C1264, "Research on Multi-Granularity Feature Fusion Multimodal Abstractive Summarization Based on Contrastive Learning").

References

[1] Li, H., Zhu, J., Liu, T., Zhang, J., & Zong, C. (2018). Multi-modal sentence summarization with modality attention and image filtering. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (pp. 4152-4158).

[2] Jangra, A., Mukherjee, S., Jatowt, A., Saha, S., & Hasanuzzaman, M. (2023). A survey on multi-modal summarization. ACM Computing Surveys, 55(13s), Article 296, 1-36.

[3] Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 379-389).

[4] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 1073-1083).

[5] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871-7880).

[6] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the 37th International Conference on Machine Learning (pp. 11328-11339).

[7] Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., & Zhou, M. (2020). ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2401-2410).

[8] Yu, T., Dai, W., Liu, Z., & Fung, P. (2021). Vision guided generative pre-trained language models for multimodal abstractive summarization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3995-4007).

[9] Xiao, M., Zhu, J., Lin, H., Zhou, Y., & Zong, C. (2023). CFSum coarse-to-fine contribution network for multimodal summarization. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8538-8553).

[10] Liu, N., Sun, X., Yu, H., Zhang, W., & Xu, G. (2020). Multistage fusion with forget gate for multimodal summarization in open-domain videos. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 1834-1845).

[11] Liang, Y., Meng, F., Xu, J., Wang, J., Chen, Y., & Zhou, J. (2023). Summary-oriented vision modeling for multimodal abstractive summarization. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 2934-2951).

[12] Zhu, J., Li, H., Liu, T., Zhou, Y., Zhang, J., & Zong, C. (2018). MSMO: Multimodal summarization with multimodal output. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4154-4164).

[13] Palaskar, S., Libovicky, J., Gella, S., & Metze, F. (2019). Multimodal abstractive summarization for How2 videos. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6587-6596).

[14] Zhang, Y., Liu, Y., Wu, S., Zhang, K., Liu, X., Liu, Q., & Chen, E. (2024). Leveraging entity information for cross-modality correlation learning: The entity-guided multimodal summarization. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 9851-9862).

[15] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748-8763).

[16] Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. In Advances in Neural Information Processing Systems 34 (pp. 9694-9705).
[17] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning (pp. 12888-12900).
[18] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (pp. 1597-1607).
[19] He, B., Wang, J., Qiu, J., Bui, T., Shrivastava, A., & Wang, Z. (2023). Align and attend: Multimodal summarization with dual contrastive losses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14867-14878).
[20] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems 26 (pp. 2787-2795).
[21] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, 457-479.
[22] Nallapati, R., Zhai, F., & Zhou, B. (2017). SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (pp. 3075-3081).
[23] Clarke, J., & Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. Journal of Artificial Intelligence Research, 31, 399-429.
[24] Zhou, Q., Yang, N., Wei, F., & Zhou, M. (2017). Selective encoding for abstractive sentence summarization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 1095-1104).
[25] UzZaman, N., Bigham, J. P., & Allen, J. F. (2011). Multimodal summarization of complex sentences. In Proceedings of the 16th International Conference on Intelligent User Interfaces (pp. 43-52).
[26] Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (pp. 4444-4451).
[27] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (pp. 74-81).
[28] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311-318).
[29] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In Proceedings of the Eighth International Conference on Learning Representations.