A Sparse-Gated Cross-Modal Alignment Framework for Sewer Defect Image-Text Retrieval

Chuanxi Zhu*, Kehe Wu
School of Control & Computer Science North China Electric Power University Beijing, China
Corresponding Author: Chuanxi Zhu (zcx666@ncepu.edu.cn)

Abstract: We present SGC-Align, a Sparse-Gated Cross-Modal Alignment framework for pipeline defect image-text retrieval. SGC-Align replaces dense pairwise similarity with a learnable Sparse-Gated Hybrid Similarity Mask (HSM) that selectively filters and reweights cross-modal interaction channels—suppressing noise, amplifying semantically salient signals, and enabling interpretable mask visualizations. To capture fine-grained correspondences, we introduce Multidirectional Cyclic Cross-Modal Attention (MCCA), a multi-round, bidirectional attention mechanism that iteratively refines local-to-text alignments. We further propose a sparsity-oriented alignment loss (LHSM) jointly trained with the mask to encourage sparse and semantically coherent matches. Extensive experiments on pipeline inspection retrieval settings demonstrate substantial improvements in Recall@K and mAP against strong baselines, as well as reduced computation and memory overhead due to the sparse-first design.

Keywords: Culture-tourism integration; Hainan Free Trade Port; Tourism English; Talent training; Industry-education integration

1. Introduction

Recent advances in vision-language pretraining and contrastive learning have led to dramatic improvements in image-text retrieval and related tasks [1,2,3,4]. Foundation models such as CLIP and ALIGN demonstrate that largescale image text supervision yields transferable cross-modal embeddings [1,2], while finer grained contrastive strategies (e.g., FILIP) and caption-oriented models (CoCa, BLIP/BLIP-2, ALBEF) push alignment quality further for downstream retrieval and understanding [3,5,6,7]. Despite these gains, industrial inspection scenarios such as pipeline defect image-text retrieval raise specific challenges — defects are often highly local, vary considerably in scale, and textual descriptions are terse, imprecise, or spatially vague (e.g., 'middle-left, large'), while images include strong background clutter and instrumentation markings that induce spurious correlations. Under such conditions, dense pairwise similarity computations are prone to propagate noise, miss fine-grained defect cues, and incur heavy computation [8].

A substantial body of work addresses related subproblems along complementary directions. Improving pretraining and transfer methods aims to give stronger initial representations and more sample-efficient adaptation (e.g., ALBEFBLIP, BLIP-2) [7,5,6,3,4,9]. Another major strand focuses on richer cross-modal interactions: multi-round attention and cross-attention architectures such as LXMERT, UNITER, Visual BERT refine fine-grained correspondences but come with increased

interaction cost [10,11,12]. Detailed region level modeling and bottom-up attention have proven effective for phrase-region matching and captioning tasks, while early image-text alignment works (e.g., VSE++, SCAN, order embeddings) laid foundations for embedding-based retrieval and hard negative mining [8].

Complementary methodological advances in sparsity, gating, and long-context modeling suggest new pathways to efficient and robust cross-modal alignment. Sparse and long sequence transformers (Sparse Transformer, Big Bird, Long former) and gating/attention variants demonstrate that selective information flow can reduce redundancy and improve robustness in noisy or long-range scenarios [13,14]. Meanwhile, progress in contrastive visual representation learning (SimCLR, MoCo, CPC) and hierarchical/transformer vision backbones (ViT, Swin) underpin stronger unimodal features for downstream fusion [15]. Interpretability techniques (e.g., Grad-CAM) and text augmentation/generation methods (Show and Tell, Show, Attend and Tell) have also been used to make model decisions more transparent and to enrich sparse textual descriptions in specialized domains [16].

Despite these complementary advances, two gaps remain for pipeline defect retrieval. First, approaches that rely on dense interaction mechanisms suffer in noisy, small-region detection contexts and scale poorly in computation and memory; conversely, coarse global embeddings miss subtle local defect signatures [8]. Second, while sparsity and gating concepts have proven their value in sequence and vision tasks, there is limited work that integrates sparsity as a principled, learnable component of cross-modal alignment with an accompanying alignment objective and interpretability considerations [13,14,4].

To address these gaps, we propose SGC-Align: A Sparse-Gated Cross-Modal Alignment framework tailored to pipeline defect image-text retrieval. The central idea is to replace uniform dense similarity computations with a learnable, hybrid sparse gating mechanism—the Sparse-Gated Similarity Mask—that selectively filters and reweights cross-modal interaction channels. This selective masking suppresses noisy or irrelevant features while preserving and amplifying channels that encode defect-text correspondences, enabling both efficiency and interpretability.

SGC-Align consists of three components. (1) A Sparse Gated Similarity Mask that fuses local and global similarity cues into a learnable sparse interaction pattern. (2) Multidirectional Cyclic Cross-Modal Attention (MCCA), which performs multi-round bidirectional refinement between visual regions and text tokens to capture fine-grained semantic and geometric relations. (3) A sparsity-oriented alignment loss that jointly optimizes selective interaction and retrieval performance. An auxiliary Global Non-Linear Compressor (GNLC) further improves representation compactness without altering the core framework.

In summary, we introduce a sparse-gated cross-modal alignment paradigm for defect retrieval, combining learnable semantic filtering, cyclic bidirectional attention, and sparsity-aware optimization to achieve efficient, fine-grained, and interpretable image-text alignment.

2. Methodology

2.1 Problem Definition and Method Overview

We consider the problem of learning sparse, robust alignments between image and text modalities under limited and potentially noisy annotations. Our objective combines a sparsity-oriented local supervision term and a global contrastive term:

$$\min \mathcal{L} = \lambda \mathcal{L}_{HSM} + \mathcal{L}_c \quad (1)$$

where $\mathcal{L}_h$ encourages sparse, high-confidence local alignments and L_c preserves global discriminability. Figure 1 gives a high-level overview of the pipeline: (i) label-based text synthesis and token encoding; (ii) GNLC preprocessing and MCCA multi-round interactions; (iii) sparsity-oriented target construction and loss computation. Implementation details are deferred to Sec. 2.2-2.4.

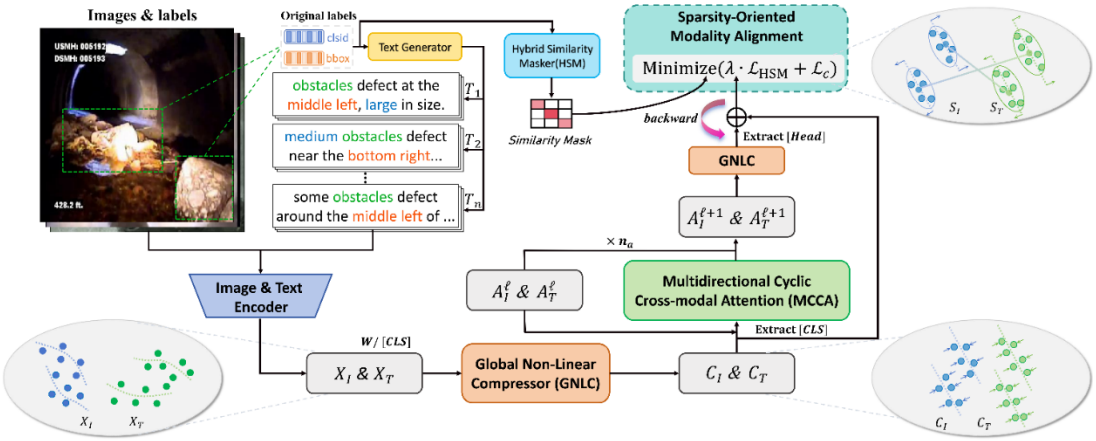

Figure 1: SGC-Align framework overview. Left: image & text inputs; Center: encoders and GNLC; Top-right: Sparse-Gated Similarity Mask generation; Bottom-right: MCCA multi-round interactions; Losses indicated by dashed arrows.

2.2 Data and Representation Construction

Text synthesis. YOLO-style labels (cls, cx, cy, w, h) are deterministically converted to short sentences by a template generator. The generator parses labels, maps class ids to names, buckets positions and sizes, fills templates and postprocesses candidates including deduplication, punctuation normalization, and truncation to at most M captions.

Token encoders. We employ transformer encoders to obtain token sequences $X_I \in \mathbb{R}^{B\times n_i\times d_x}$, $X_T \in \mathbb{R}^{B\times n_t\times d_x}$, where $X[:,0,:]$ denotes the aggregation token CLS used for retrieval.

We adopt BERT for text and Swin Transformer for vision: BERT offers bidirectional contextualization suitable for fine-grained alignment, and Swin balances local detail and global context with favorable compute/accuracy tradeoffs.

GNLC and the final embedding formation. Illustrates GNLC. We apply GNLC twice: first to obtain $C = \text{GNLC}(X)$ before MCCA, and again after MCCA to get $C' = \text{GNLC}(A)$. Denote $\text{CLS}' = C[:,0,:]$ and $Head = C'[:,0,:]$, the final sample embedding is obtained by element-wise addition:

$$F = CLS' + Head \tag{2}$$

which is used directly for the contrastive and HSM losses.

2.3 Multidirectional Cyclic Cross-Modal Attention (MCCA)

As shown in Figure 3, MCCA is an iterative, modular encoder that refines image and text token representations via bidirectional message passing. The pipeline initializes with

GNLC outputs, i.e. $A_I^0 = C_I$ and $A_T^0 = C_T$. Each MCCABlock consumes the current-layer token sequences A_I^{ℓ}, A_T^{ℓ} and produces updated sequences $A_I^{\ell+1}, A_T^{\ell+1}$. For brevity let $A^{\ell} \in \mathbb{R}^{B\times n\times d_c}$ denote a generic token sequence; $\text{MHA}(Q,K,V)$ denotes multi-head attention with query Q, key K and value V, $\text{MLP}(\cdot)$ and $\text{LN}(\cdot)$ denote feed-forward block and layer normalization, respectively.Each MCCABlock comprises a Self-Attention stage followed by a Cross-Attention stage. The Self-Attention

stage applies residual attention and a feed-forward update:

$$\tilde{A}^{\ell} = A^{\ell} + \mathrm{LN}(\mathrm{MHA}(A^{\ell}, A^{\ell}, A^{\ell})) \tag{3}$$

$$\overline{A}_T^{\ell} = \hat{A}_T^{\ell} + \mathrm{LN}\big(\mathrm{MHA}(\hat{A}_I^{\ell}, \hat{A}_T^{\ell}, \hat{A}_T^{\ell})\big) \tag{4}$$

$$A_T^{\ell+1} = \overline{A}_T^{\ell} + \mathrm{LN}\left(\mathrm{MLP}(\overline{A}_T^{\ell})\right) \tag{5}$$

$$\overline{A}_I^{\ell} = \hat{A}_I^{\ell} + \mathrm{LN}\big(\mathrm{MHA}(\hat{A}_T^{\ell}, \hat{A}_I^{\ell}, \hat{A}_I^{\ell})\big) \tag{6}$$

$$A_I^{\ell+1} = \overline{A}_I^{\ell} + \mathrm{LN}\left(\mathrm{MLP}(\overline{A}_I^{\ell})\right) \tag{7}$$

2.4 Sparsity-Oriented Modality Alignment

HSM construction HSM combines class agreement and bounding-box overlap into a sparse, row-stochastic supervisory mask S. Concretely, for a mini-batch of N samples with labels $\{y_i\}$ and boxes $\{b_i\}$ we compute:

$$S_{\mathrm{raw},ij} = \mathrm{cls\_w} \cdot \mathbf{1}[y_i = y_j] \; + \; \mathrm{bbox\_w} \cdot g\big(\mathrm{IoU}(b_i, b_j)\big) \tag{8}$$

$$g(x) = \begin{cases} x \cdot \mathbf{1}[x > \tau], & \text{(hard threshold)} \\ x \cdot \sigma((x - \tau)\alpha), & \text{(soft gating)} \end{cases} \tag{9}$$

To obtain a proper row-stochastic mask we normalize each row with numerical protection:

$$S_{i,:} = \frac{S_{\mathrm{raw},i,:}}{\sum_j S_{raw,i,j} + \varepsilon}, \text{ if } \sum_j S_{raw,i,j} \geq \varepsilon \tag{10}$$

$$S_{i,:} = e_i^{\top} \text{ (fallback to self-similarity) otherwise} \tag{11}$$

where e_i is the i-th canonical basis vector and $\varepsilon > 0$ avoids divide-by-zero. Gating/threshold choices control sparsity (few nonzero neighbors per row), and the row normalization converts those sparse neighborhoods into probabilistic soft centers.

Loss function Let $F^I, F^T \in \mathbb{R}^{N \times d}$ be the final per-sample embeddings in the current batch, obtained as $F = \mathrm{CLS} + Head$ where $\mathrm{CLS} = C[:,0,:]$ and $Head = C'[:,0,:]$. We jointly optimize a symmetric contrastive loss and an HSM loss guided by the row-stochastic mask S. Contrastive (InfoNCE-style) objectives are used for stable global discrimination

For contrastive learning define pairwise similarities:

$$s_{ij} = \frac{(F_i^I)^{\top} F_j^T}{\tau} \tag{12}$$

The image-to-text and text-to-image cross-entropy terms are:

$$L_{\mathrm{img2txt}}(i) = -\log \frac{\exp(s_{ii})}{\sum_j \exp\,(S_{ij})} \tag{13}$$

$$L_{\mathrm{txt2img}}(i) = -\log \frac{\exp(s_{ii})}{\sum_j \exp\,(S_{ji})} \tag{14}$$

Table 1: Retrieval Results on Joint-SDPR. Image-query-Text: Use Image as Query to Retrieve Captions; Text-query-Image: Use Text as Query to Retrieve Images.

Method (Loss)	Backbone (vision/text)	Image-query-Text R@1/R@5/R@10	Text-query-Image R@1/R@5/R@10	mR
VSE0 (L_t)	ResNet-50 / GRU	10.13 / 31.79 / 41.01	16.37 / 32.21 / 41.29	28.80
CAMP (L_t)	RoI Transformer / GRU	10.30 / 27.76 / 43.07	15.91 / 33.17 / 45.31	29.25
CAMERA (L_t)	RoI Transformer / BERT	10.15 / 24.85 / 38.91	14.98 / 33.39 / 43.93	27.70
LMW-MCR (L_t)	CNN Blocks / CNN Blocks	7.31 / 23.79 / 35.63	17.45 / 36.69 / 48.90	28.30
AMFMN (L_t)	ResNet-18 / GRU	11.57 / 27.97 / 38.57	15.28 / 35.63 / 49.87	29.82
GaLR (L_t)	ResNet-18 + ppyolo / GRU	14.63 / 37.72 / 55.49	17.56 / 40.84 / 52.36	36.43
KCR (L_t)	ResNet-101 / BERT	13.21 / 35.32 / 52.88	20.23 / 47.02 / 60.83	38.25
SWAN (L_t)	ResNet-50 / GRU	16.46 / 38.25 / 55.18	20.82 / 46.66 / 60.90	39.71
VSE1 (L_c)	ViT / BERT	20.12 / 43.28 / 58.57	19.93 / 40.80 / 54.96	39.61
VSE2 (L_t)	Swin Transformer / BERT	17.69 / 41.19 / 61.33	24.11 / 46.30 / 61.07	41.95
VSE2 (L_c)	Swin Transformer / BERT	23.37 / 47.61 / 65.77	22.87 / 49.53 / 62.33	45.25
SGC-Align (L_c +L_{HSM})	Swin Transformer / BERT	22.07 / 52.31 / 70.93	26.48 / 51.58 / 63.92	47.88

The global contrastive loss is the batch average:

$$\mathcal{L}_c = \frac{1}{2N}\sum_i L_{\text{img2txt}}(i) + L_{\text{txt2img}}(i) \tag{15}$$

The HSM loss constructs sparse, label/bbox-aware soft centers from F. Let:

$$\bar{C}^T = S\,F^T, \qquad \bar{C}^I = S\,F^I \tag{16}$$

Define center similarities (example for image-to-text):

$$s^*_{i,j}(\text{img2txt}) = \frac{(F_i^I)^\top \bar{C}_j^T}{\tau_2} \tag{17}$$

Per-sample cross-entropy terms are formed analogously to the contrastive terms and averaged to give LHSM. Because S is sparse by design (gating/thresholding plus row normalization), LHSM provides localized, multi-positive supervision that emphasizes semantically and spatially consistent neighbors and mitigates noisy annotations.

We optimize the weighted combination given in Eq. (1), where λ trades off sparse local supervision and global discrimination. Practical notes: ensure embedding normalization when using cosine similarities, include a small ε/row fallback for S, and aggregate features across replicas (all gather) in distributed training.

3. Experiments

3.1 Datasets and Evaluation Metrics

We use the Joint-SDPR dataset, merged from Storm Drain and Pipe-root, comprising eight defect classes and 2,849 annotated instances. Images are at 416 × 416 resolution. Text descriptions are generated from bounding boxes and class labels via the template generator and tokenized per model configuration.

We report R@K for K ∈ {1, 5, 10} (the percentage of queries with at least one ground-truth in the top-K) and mean recall mR (the average of R@1, R@5 and R@10). For multi ground-truth queries, a hit occurs if any ground truth appears within the top-K. Reported metrics are averaged over multiple

runs.

3.2 Implementation Details

We use a Swin Transformer visual encoder (Swin-T and a BERT text encoder (BERT-base) initialized from the official release. Both encoders produce 768-dimensional token features, which are linearly projected to 512-dimensional embeddings. Self and cross-attention use 8 heads with dropout 0.2, and the MCCA stack comprises 3 MCCABlocks. The temperature for both the contrastive and attribution losses is set to 0.07 and the factor λ is 1. Training was performed on two NVIDIA RTX 4090 GPU with batch size 64 using the AdamW optimizer (learning rate 6e−5, weight decay 0.01).

3.3 Performance Comparison

We compare SGC-Align to representative prior methods from three categories: global joint embedding methods such as VSE0; region-based approaches that model local regions or multi-scale cues, including CAMP, LMW-MCR, CAMERA, AMFMN, GaLR, KCR and SWAN; and contrastive pretrained baselines VSE1 and VSE2 that use strong vision and text backbones. For clarity, VSE0-VSE2 denote variants of the VSE that differ in their vision and text backbones. We evaluated the retrieval results of Image-query-Text (I2T) and Text-query-Image (T2I). Table 1 lists each method, s vision/text backbone and primary loss where Lt denotes triplet loss, Lc denotes contrastive (InfoNCE) loss and LHSM denotes the proposed HSM loss.

SGC-Align achieves the highest mean recall, mR 47.88%, outperforming the strongest baseline VSE2 trained with Lc by 2.63%, from 45.25% to 47.88%. The most notable single improvement is T2I R@1, rising from 22.87% to 26.48%. Top-K metrics R@5 and R@10 also improve consistently, indicating better multi-positive and top-K coverage. I2T R@1 is slightly lower than VSE2 (22.07% vs 23.37%), suggesting a minor trade-off between immediate top-1 precision and overall robustness across multiple relevant targets. Overall, the consistent gains in Top-K recall and mean recall show that SGC-Align delivers stronger and more robust retrieval performance across different query types.

Figure 4 shows qualitative Top-5 retrieval examples for representative queries. Qualitatively, I2T results are largely correct with Top-5 lists often containing multiple valid captions for the same defect. For T2I retrieval the top-3 returned images frequently include correct instances, indicating strong semantic matching. Remaining failures are typically due to errors in the described spatial position or relative size (for example, 'middle left' vs 'middle right' or 'large' vs 'medium') rather than misclassification of defect type, suggesting room for improvement in precise localization and scale matching.

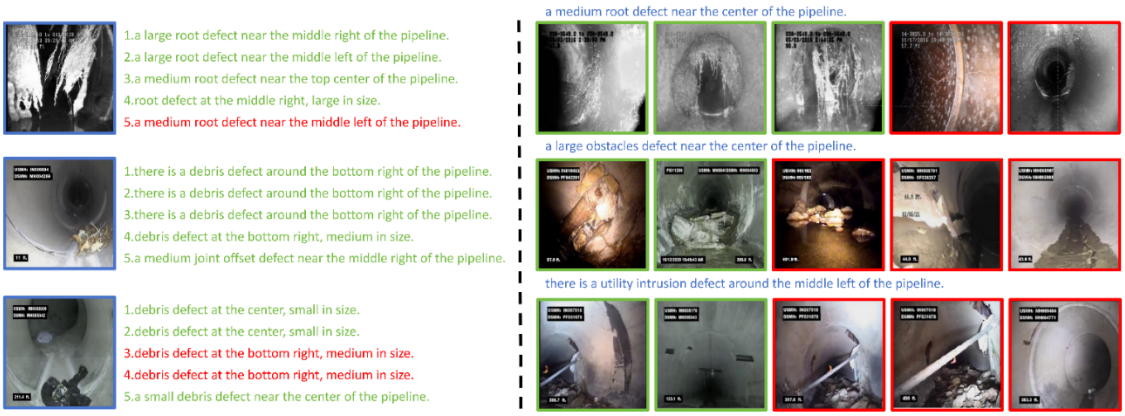

Figure 4: Top-5 retrieval visualization for each query. The left panel shows text candidates retrieved for image queries; the right panel shows image candidates retrieved for text queries. Blue indicates the query item, green indicates correct matches, and red indicates incorrect or irrelevant candidates (Top-5).

Table 2: Results of ablation experiments with different configurations on the Joint-SDPR dataset.

Method	**Setting** (MCCA / L_{HSM})	**Image-query-Text** R@1 / R@5 / R@10	**Text-query-Image** R@1 / R@5 / R@10	**mR**
baseline	- / -	18.62 / 42.17 / 58.47	17.18 / 40.36 / 51.12	37.99
+ MCCA	n_a = 3 / -	20.87 / 49.43 / 68.21	23.74 / 49.61 / 62.37	45.64
+ L_{HSM}	- / cls-only	19.53 / 47.92 / 64.71	20.91 / 46.77 / 58.14	43.05
+ L_{HSM}	- / bbox-only	19.41 / 47.68 / 64.56	20.61 / 46.51 / 58.33	42.79
+ L_{HSM}	- / cls + bbox	19.77 / 48.12 / 65.43	21.05 / 47.08 / 59.62	43.51
+ MCCA + L_{HSM}	n_a = 3 / cls-only	21.42 / 51.05 / 69.10	24.55 / 50.60 / 62.35	46.51
+ MCCA + L_{HSM}	n_a = 3 / bbox-only	21.18 / 50.62 / 69.05	24.12 / 50.02 / 63.50	46.42
+ MCCA + L_{HSM}	n_a = 3 / cls + bbox	22.07 / 52.31 / 70.93	26.48 / 51.58 / 63.92	47.88

3.4 Ablation Study

We systematically evaluate the individual and combined effects of L_{HSM} and MCCA. Here n_a denotes the number of MCCA blocks. In the HSM configuration, 'cls-only' enables only the class weight, 'bbox-only' enables only the bounding-box weight, and 'cls + bbox' enables both and combines them linearly. The baseline refers to the model configuration in which both MCCA and L_{HSM} are disabled. All reported retrieval scores are given as percentages.

MCCA contribution analysis. Introducing MCCA with the number of interaction rounds n_a set to 3 yields substantial improvements relative to the baseline: mean recall increases from 37.99% to 45.64%, an improvement of 7.65%. I2T R@1 improves from 18.62% to 20.87% (+2.25%), while T2I R@1 rises from 17.18% to 23.74% (+6.56%). R@5 and R@10 exhibit similar gains, indicating enhanced top-K coverage when multi-round, bidirectional interactions are enabled. These results indicate that MCCA strengthens top-1 accuracy and improves robustness in top-K retrieval, particularly under text-query scenarios.

HSM Loss contribution analysis Applying L_{HSM} (cls + bbox) alone increases mean recall from 37.99% to 43.51% (+5.52%), with the largest effects observed on R@5 and R@10 (For instance, I2T R@5 increases by 5.95%). Single signal variants L_{HSM} (class-only) or L_{HSM} (bbox-only) yield intermediate gains. When L_{HSM} (class-only or bbox-only) is combined with MCCA, further improvements are obtained, and the full combination attains the best performance (mR reaches 47.88%). Overall, LHSM and MCCA provide complementary benefits: L_{HSM} primarily enhances multi-positive and top-K coverage while MCCA contributes stronger top-1 improvements, and their combination yields the largest net gain.

3.5 Parameter Study

We analyze two hyperparameters that materially affect performance: the number of MCCA rounds n_a and the HSM weight coefficient λ. The experiments use Swin-T/BERT with full HSM unless noted. Figure 5 and Tables 4-3 summarize results; below we separate the analysis into two focused aspects.

Sensitivity to number of MCCA rounds (n_a). Table 4 and the top-left panel of Figure 5 show that

increasing n_a yields substantial gains from n_a=1 to n_a=3 (mR: 44.10→47.88). The improvements manifest across metrics but are most pronounced in R@5/R@10, indicating stronger top-K coverage and improved multi-positive retrieval—consistent with MCCA's role in iteratively refining local-text correspondences. Gains diminish beyond n_a=3 (negligible changes for $n_a \geqslant 4$) while compute and memory scale roughly linearly with the number of rounds. Practical tradeoff: use n_a=3 for best accuracy-cost balance; use n_a=2 if resources are constrained but you still need much of the benefit.

Effect of HSM weight coefficient (λ) and embedding geometry. Table 3 and the top-right panel of Figure 5 show a peaked response with an optimum at λ=1 (mR = 47.88).

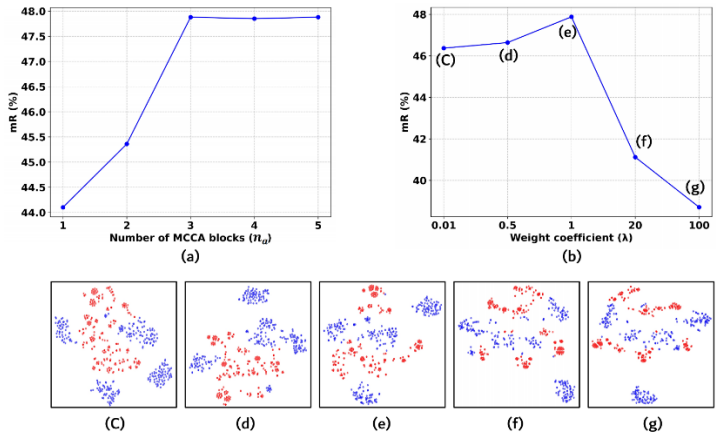

Figure 5: Parameter study and corresponding embedding visualizations. Top-left: mean recall (mR) as a function of the number of MCCA blocks (n_a). Top-right: mR as a function of the weight coefficient (λ); labeled points (C)–(G) on this plot correspond to the five embedding visualizations shown on the bottom row. Bottom row (left to right): embedding visualizations corresponding to points (C)–(G) in the top-right plot.

Table 3: Effect of weight coefficient λ on retrieval performance (sorted by λ ascending).

λ	**Image-query-Text** R@1 / R@5 / R@10	**Text-query-Image** R@1 / R@5 / R@10	**mR**
0.01	19.25 / 50.00 / 68.66	26.87 / 49.25 / 64.18	46.37
0.5	20.00 / 50.00 / 68.06	23.88 / 50.75 / 67.16	46.64
1	22.07 / 52.31 / 70.93	26.48 / 51.58 / 63.92	47.88
20	15.52 / 45.82 / 62.99	21.64 / 44.03 / 56.72	41.12
100	12.99 / 40.75 / 59.10	22.39 / 41.04 / 55.97	38.71

Table 4: Effect of number of MCCA rounds n_a (full HSM enabled).

na	**Image-query-Text** R@1 / R@5 / R@10	**Text-query-Image** R@1 / R@5 / R@10	**mR**
1	20.95 / 48.62 / 66.20	21.48 / 47.33 / 60.02	44.10
2	21.30 / 49.12 / 67.05	22.48 / 48.80 / 63.42	45.36
3	22.07 / 52.31 / 70.93	26.48 / 51.58 / 63.92	47.88
4	22.03 / 52.14 / 70.98	26.43 / 51.56 / 63.97	47.85
5	22.09 / 52.22 / 70.95	26.49 / 51.59 / 63.94	47.88

Small HSM weights (e.g., λ=0.01,0.5) help structure embeddings without harming global discrimination (mR$\approx$46.4-46.6), while very large weights (e.g., λ=20,100) degrade performance substantially. The bottom embedding visualizations (points (C)−(G)) explain why: (C) and (D) (small λ) show reasonably separated clusters with some intra-class spread; (E) (λ=1) achieves compact intra-class clusters and clear inter-class separation; (F) and (G) (large λ) show fragmented clusters and spurious substructures that reduce global retrieval quality. Interpretation: HSM encourages desirable local grouping (multi-positive supervision) but overemphasis causes intra-class overfitting and loss of global margins.

4. Conclusion

We presented SGC-Align, a Sparse-Gated Cross-Modal Alignment framework tailored for pipeline defect image-text retrieval. The method replaces dense pairwise similarity with a learnable Hybrid Similarity Mask that selectively filters and reweights cross-modal channels, and employs Multidirectional Cyclic Cross-Modal Attention to iteratively refine fine-grained local-text correspondences. A sparsity-oriented alignment loss jointly trains the mask and interaction modules to encourage semantically coherent, sparse matches and to enable interpretable mask visualizations.

Extensive experiments on Joint-SDPR demonstrate consistent improvements over strong baselines, achieving the best mean recall (mR = 47.88) and substantial gains in multi positive and top-K coverage. Ablations show that MCCA primarily improves fine-grained R@1 performance while HSM enhances multi-ground-truth robustness; combining both yields complementary benefits. The sparse-first design also reduces computation and memory footprint and improves robustness under limited-data and domain-shift conditions.

Future work includes tighter integration with large-scale vision-language pretraining to improve zero-shot and cross domain generalization, exploring self-supervised or learned mask priors to reduce annotation dependency, and extending the framework to open-vocabulary and multi-modal industrial inspection tasks.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, Q. V. Pham, H. W. Lee, S. Sanyal, J. Maynez, A. Kolesnikov, et al., Scaling up visual and vision-language representation learning with noisy text supervision.

[3] S. Yao, et al., Coca: Contrastive captioners are image-text foundation models.

[4] X. Zhai, et al., Slip: Self-supervision meets language-image pre-training.

[5] J. Li, et al., Blip: Bootstrapping language-image pre-training.

[6] J. Li, et al., Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

[7] L. Li, X. Liang, et al., Albef: Align before fuse.

[8] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, Vse++: Improving visual-semantic embeddings with hard negatives, in: Proceedings of the British Machine Vision Conference (BMVC), 2018.

[9] H. Zhou, et al., Clip-adapter: Parameter-efficient transfer learning for vision-language models.

[10] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, in: EMNLP-IJCNLP, 2019.

[11] Y.-C. Chen, L. Li, L. Yu, A. Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text

representation learning.

[12] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic Visio linguistic representations for vision and language tasks, in: NeurIPS, 2019.

[13] R. Child, S. Gray, A. Radford, I. Sutskever, Generating long sequences with sparse transformers.

[14] M. Zaheer, et al., Big bird: Transformers for longer sequences, in: NeurIPS, 2020, pp. 17283-17297.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626.