PromptHash: Affinity-Prompted Collaborative Cross-Modal Learning for Adaptive Hashing Retrieval

Qiang Zou, Shuli Cheng and Jiayi Chen

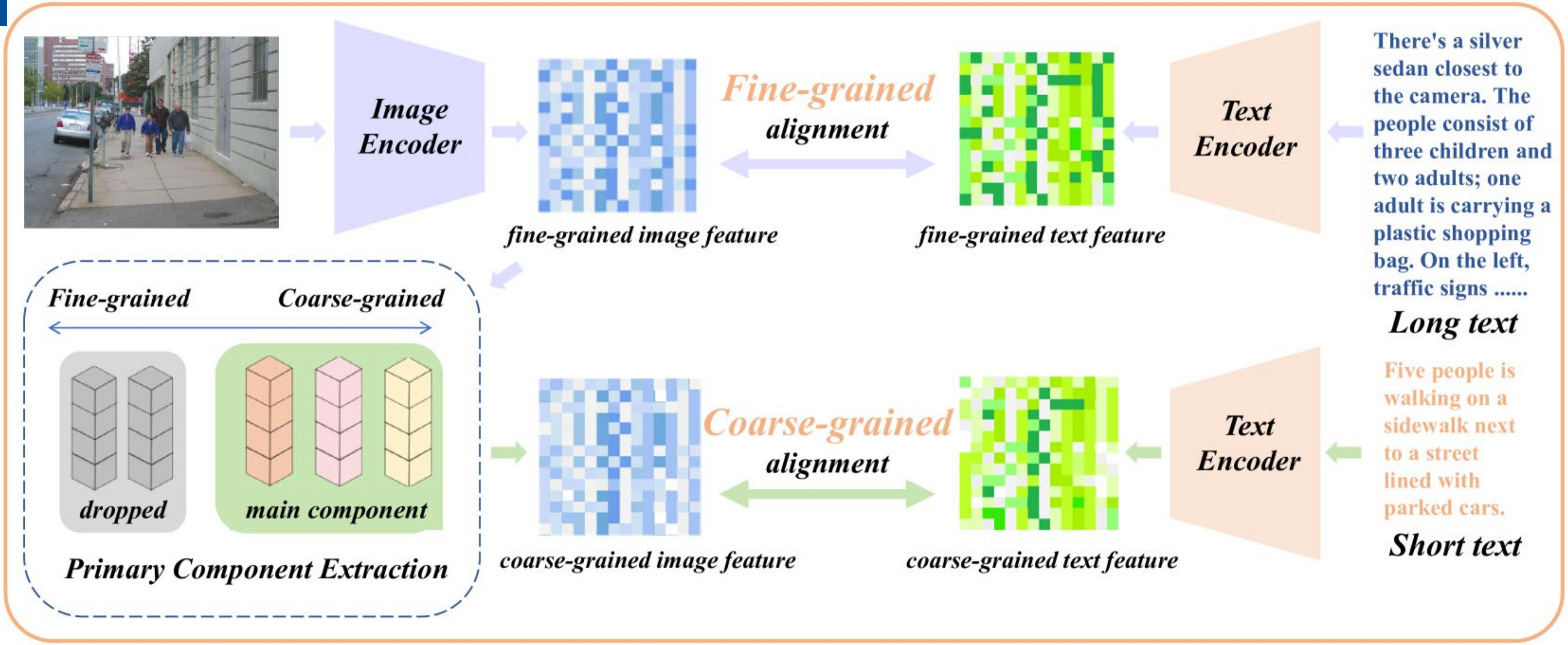
School of Computer Science and Technology, Xinjiang University

Ürümqi 830046, China

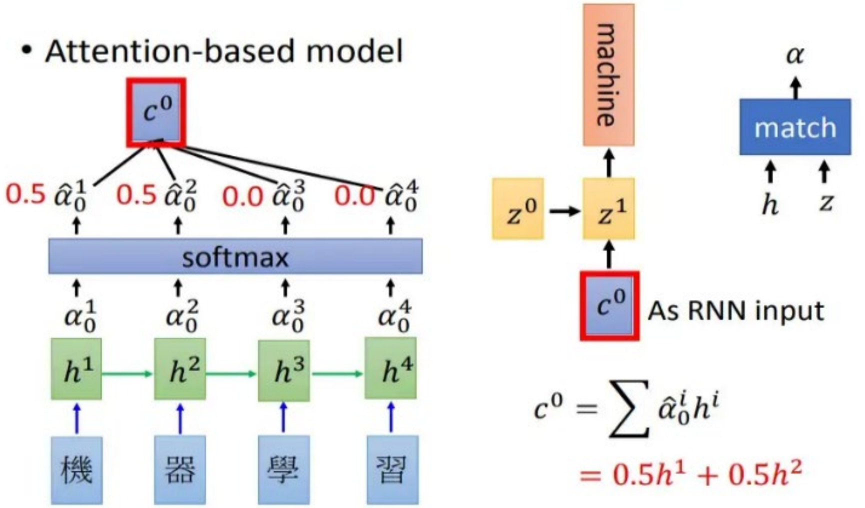
Motivation & Problem Statement

the challenge of truncated text semantics

For example, in the commonly used cross-modal hashing dataset MS-COCO, text lengths can range from 169 to 625 characters, far exceeding the maximum input limit of 77 characters for the CLIP text encoder. This results in severe semantic truncation. Therefore, mitigating this issue and supplementing missing contextual semantics has become a major challenge in the research community.



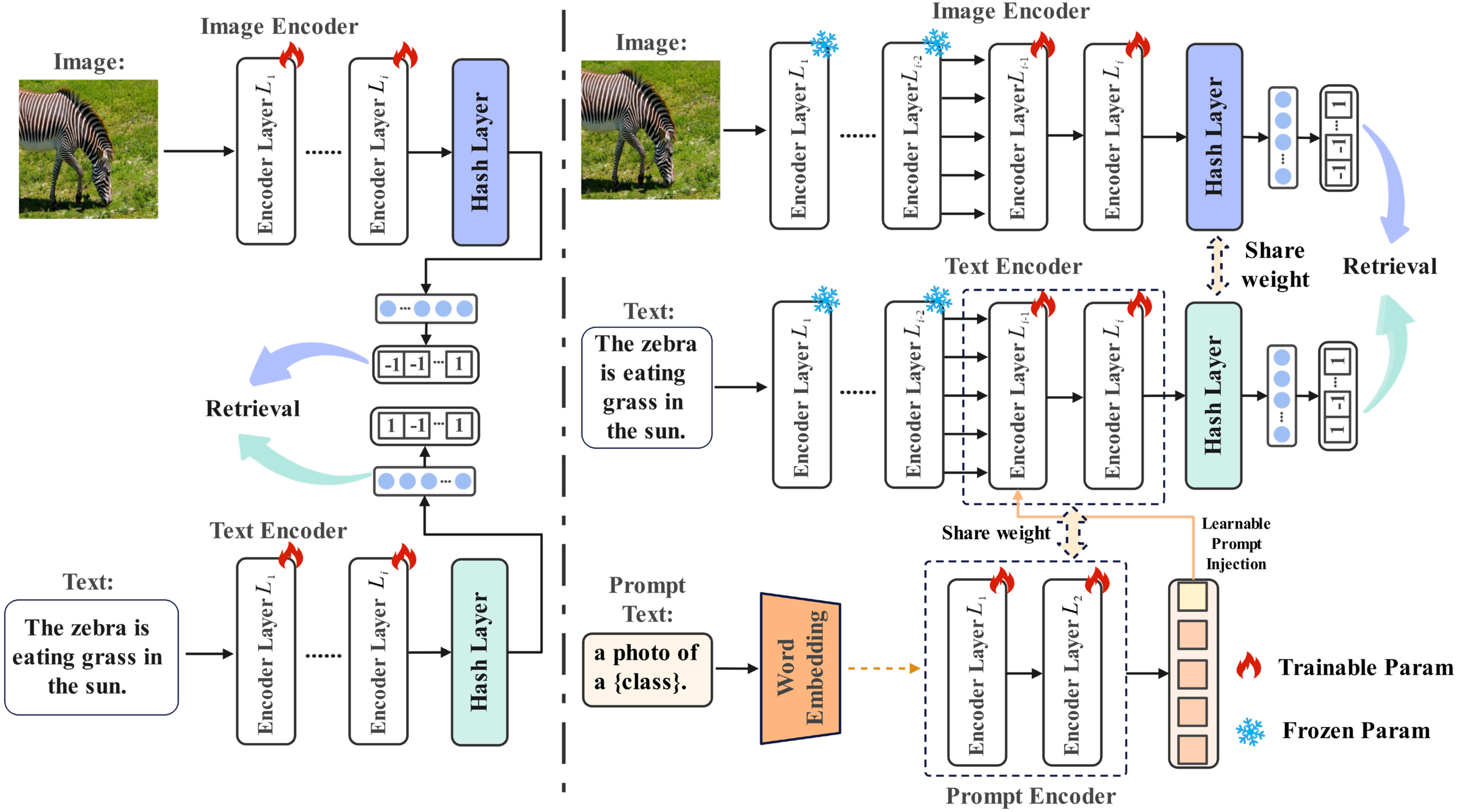
Motivation & Problem Statement



Semantic loss and contextual redundancy

Existing cross-modal hashing methods utilize contrastive learning to align modalities, but benchmark datasets like MIRFLICKR-25K, NUS-WIDE, and MS-COCO still suffer from context loss and semantic redundancy in text representations. For example, MIRFLICKR-25K and NUS-WIDE concatenate multiple tags without context, while MS-COCO merges multiple captions, leading to redundant information. These issues result in suboptimal hash codes and reduced retrieval performance.

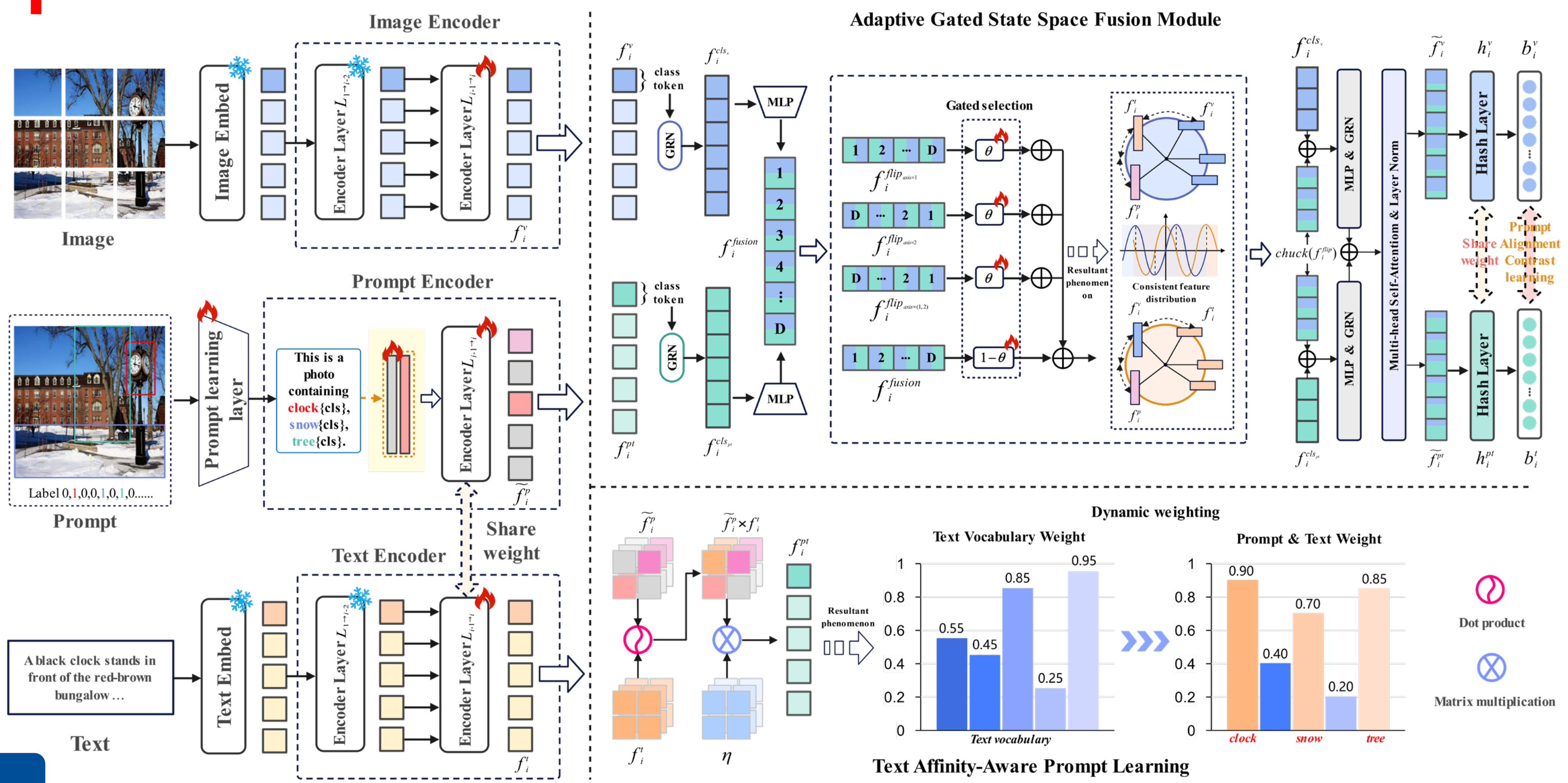
Related Work



(a) Previous method

(b) Ours method

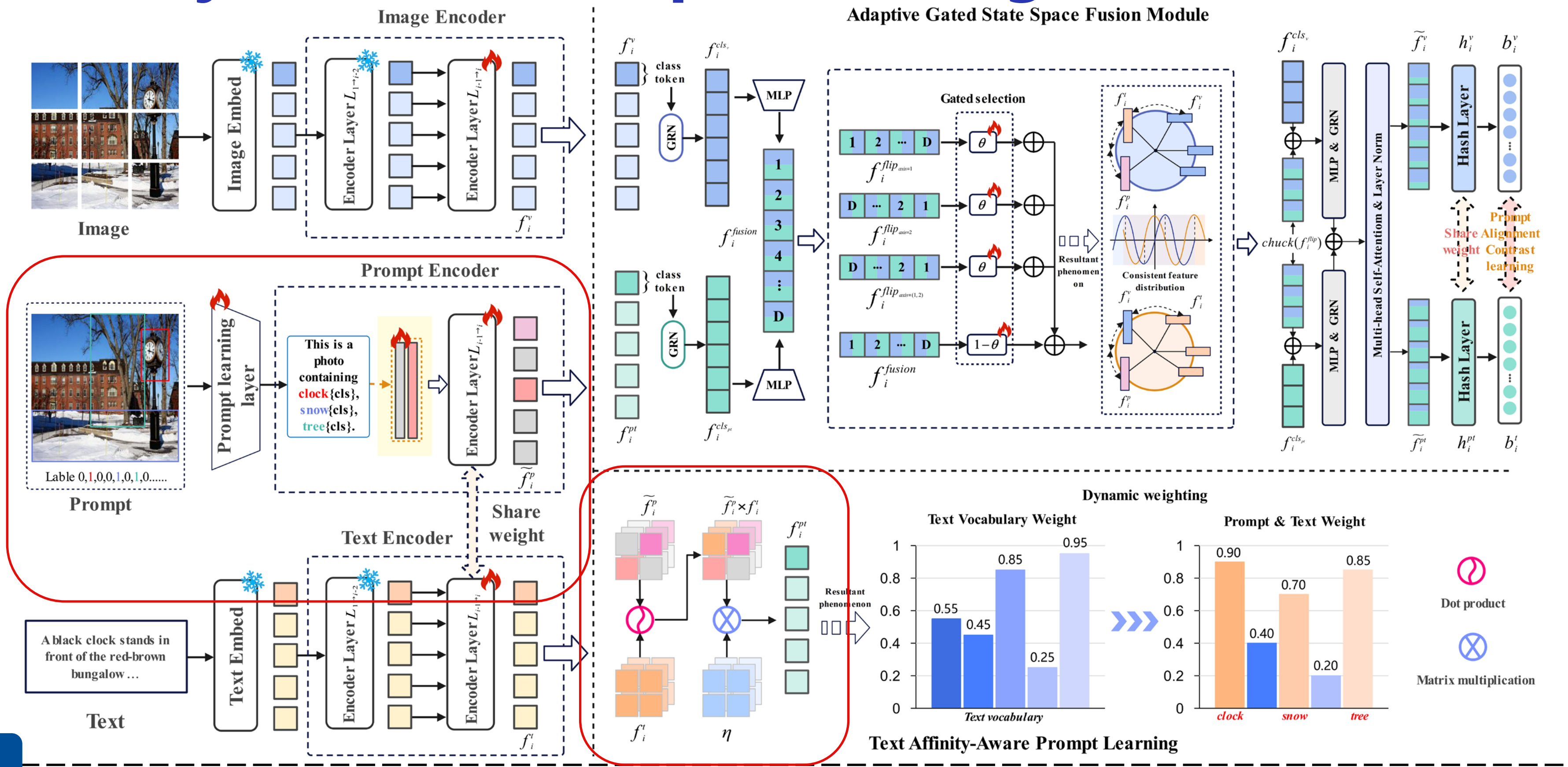
PromptHash



Research

Mainstream pre-trained models like CLIP often truncate long texts, causing semantic loss and suboptimal hash codes, while some datasets suffer from context deficiency and redundancy. To address this, we introduce a learnable Text Affinity-Aware Prompt (TAAP) to highlight retrieval-relevant semantics and mitigate truncation. Additionally, an Adaptive Gated Selection Fusion (AGSF) module adaptively fuses multimodal features and filters redundant information, further improving cross-mod al retrieval.

Text Affinity-Aware Prompt Learning (TAAP)

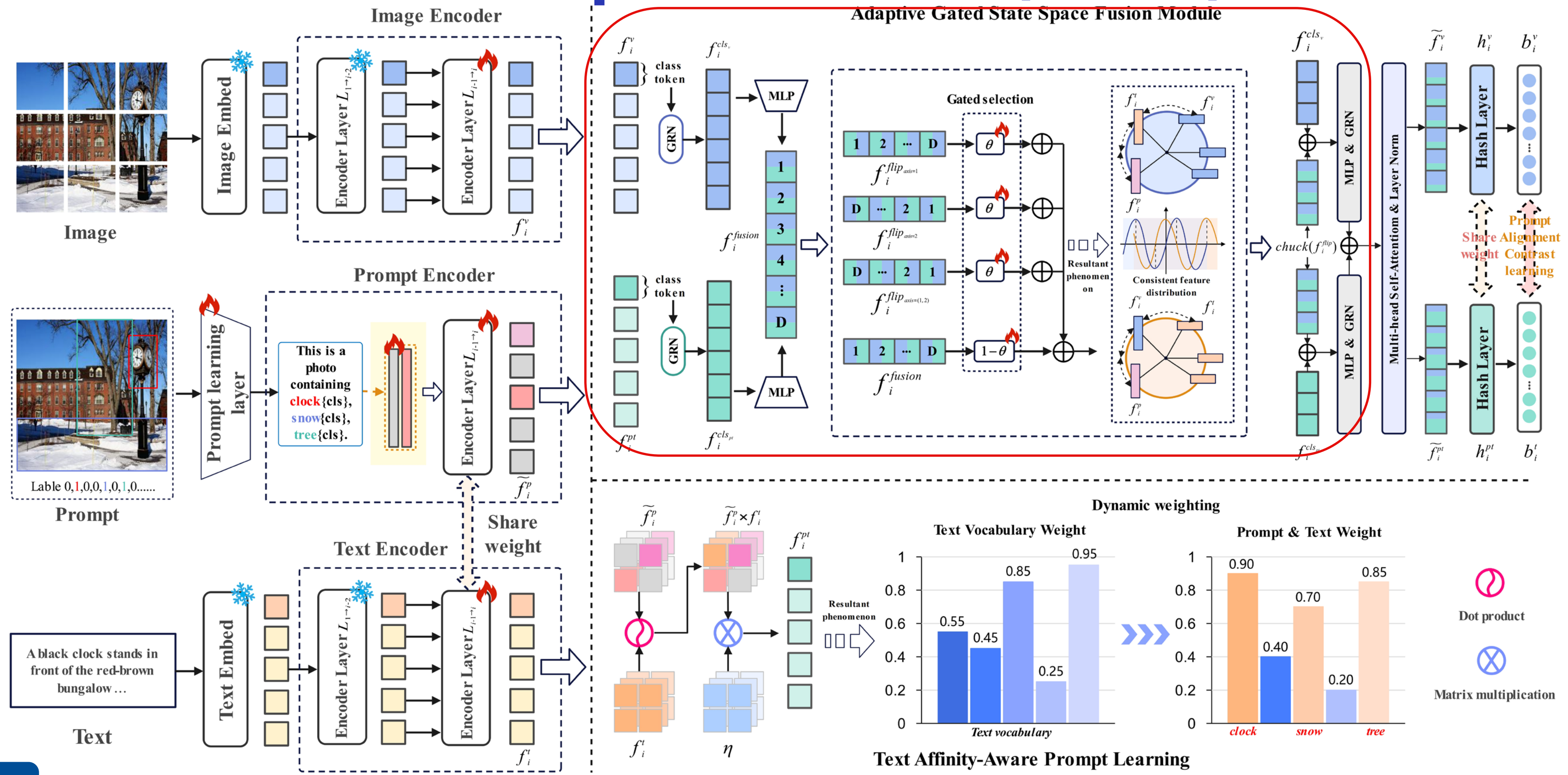


Text Affinity-Aware Prompt Learning

Research

To mitigate semantic truncation and context loss, we introduce a learnable affinity text prompt, where all tokens except the class name are learnable. The prompt is trained using the last two Transformer layers of the CLIP text encoder and adaptively fused with original text features via an adapter, thus highlighting retrieval-relevant semantics and improving retrieval performance without extra visual prompts.

Adaptive Gated State Space Fusion (AGSF)



Research

To address cross-modal redundancy, we propose an Adaptive Gated Selection Fusion (AGSF) module that combines SSM and Transformer strengths to adaptively fuse features, emphasizing retrieval-relevant semantics and filtering out redundant or negative information.

Experimental Results--Ablation Study

	Methods	MIRFLICKR-25K			NUS-WIDE			MS COCO		
		16bit	32bit	64bit	16bit	32bit	64bit	16bit	32bit	64bit
I2T	baseline	0.8444	0.8590	0.8666	0.7228	0.7377	0.7488	0.7099	0.7594	0.7793
	w/o (PACL+AGSF)	0.8767	0.8855	0.8908	0.8278	0.8555	0.8656	0.7222	0.7910	0.8207
	w/o (TAAP+PACL)	0.9139	0.9269	0.9329	0.7668	0.7838	0.7893	0.7117	0.7687	0.7933
	w/o AGSF	0.9346	0.9512	0.9593	0.8377	0.8584	0.8709	0.7424	0.8111	0.8444
	w/o PACL	0.9720	0.9859	0.9924	0.9013	0.9450	0.9639	0.7246	0.8228	0.8725
	PromptHash(Ours)	**0.9818**	**0.9960**	**0.9995**	**0.9313**	**0.9759**	**0.9931**	**0.7673**	**0.8782**	**0.9263**
T2I	baseline	0.8383	0.8488	0.8524	0.7303	0.7477	0.7588	0.7201	0.7595	0.7885
	w/o (PACL+AGSF)	0.8558	0.8641	0.8704	0.7996	0.8272	0.8338	0.7205	0.7804	0.8103
	w/o (TAAP+PACL)	0.9140	0.9276	0.9331	0.7696	0.7841	0.7902	0.7141	0.7755	0.7973
	w/o AGSF	0.9420	0.9591	0.9670	0.8157	0.8313	0.8460	0.7509	0.8092	0.8414
	w/o PACL	0.9720	0.9859	0.9925	0.9060	0.9517	0.9676	0.7227	0.8272	0.8731
	PromptHash(Ours)	**0.9816**	**0.9955**	**0.9995**	**0.9381**	**0.9761**	**0.9934**	**0.7667**	**0.8790**	**0.9283**

Research

The table shows ablation results, indicating that the proposed TAAP module effectively mitigates text semantic truncation by adaptively weighting and preserving retrieval-relevant semantic information while suppressing irrelevant features. Additionally, the fusion module selects beneficial semantics from both modalities and filters out redundant contextual information. Aligning global and local prompt tokens further optimizes semantic representation, leading to high-quality hash codes.

Experimental Results – Comparison on NUS-WIDE

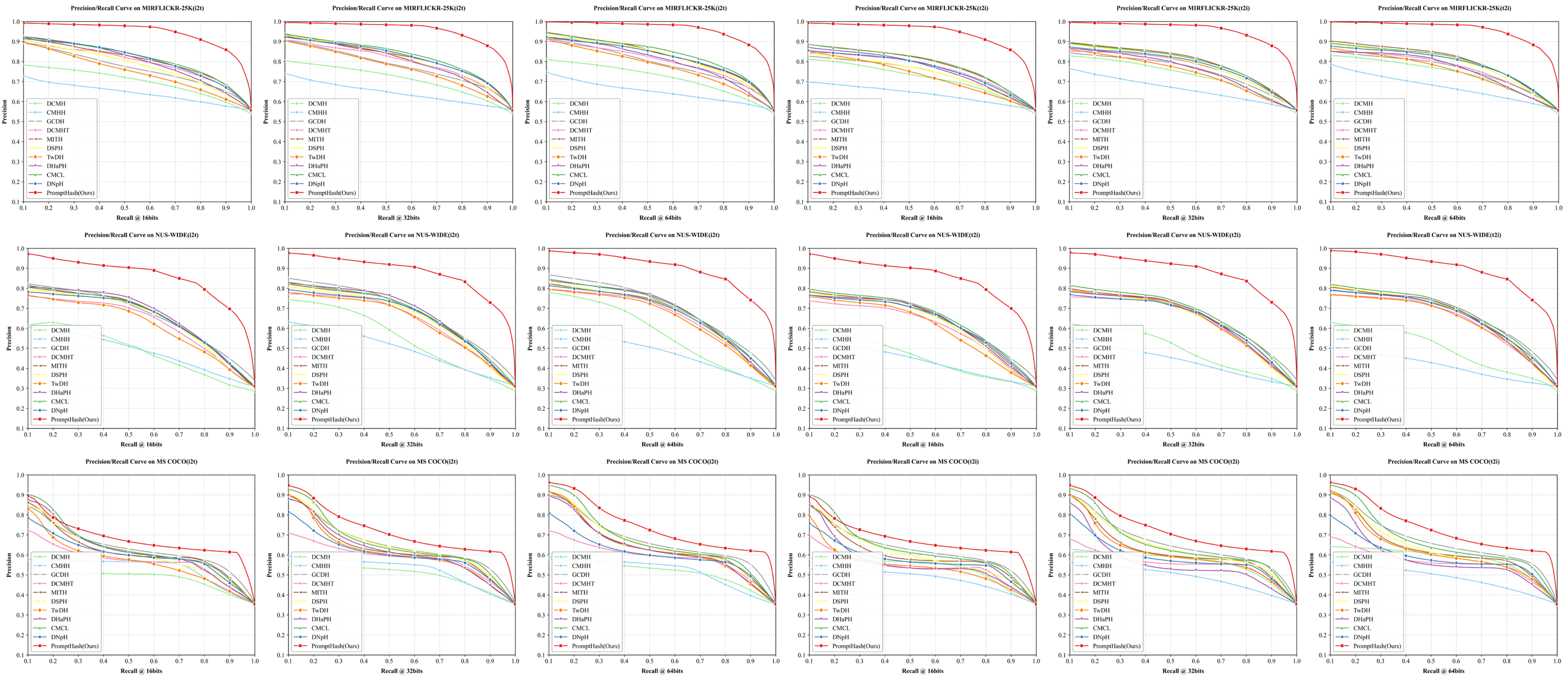
Methods	I2T				T2I			
	16bit	32bit	64bit	Avg	16bit	32bit	64bit	Avg
DCMH	0.5238	0.5995	0.6195	0.5809	0.544	0.5901	0.5956	0.5766
CMHH	0.5312	0.5476	0.5299	0.5362	0.4826	0.4868	0.4711	0.4802
GCDH	0.7142	0.7367	0.7498	0.7336	0.7215	0.7423	0.7534	0.7391
DCHMT	0.6832	0.6892	0.7025	0.6916	0.692	0.7081	0.7208	0.707
MITH	0.7062	0.718	0.7186	0.7143	0.7122	0.7281	0.7335	0.7246
DSPH	0.6953	0.7031	0.7161	0.7048	0.7028	0.7165	0.7329	0.7174
TwDH	0.6649	0.6855	0.6933	0.6812	0.6719	0.7126	0.7153	0.6999
DNpH	0.7135	0.7169	0.7247	0.7184	0.7222	0.7265	0.7313	0.7267
DHaPH	0.7215	0.7333	0.741	0.7319	0.7203	0.7284	0.7388	0.7292
CMCL	0.7154	0.7314	0.744	0.7303	0.729	0.7438	0.7511	0.7413
VTPH	<u>0.7733</u>	<u>0.7870</u>	<u>0.7936</u>	<u>0.7846</u>	<u>0.7708</u>	<u>0.7840</u>	<u>0.7933</u>	<u>0.7827</u>
PromptHash	**0.9313**	**0.9759**	**0.9931**	**0.9668**	**0.9381**	**0.9761**	**0.9934**	**0.9692**

Methods	I2T			T2I		
	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
baseline	0.7228	0.7377	0.7488	0.7303	0.7477	0.7588
w/o (PACL + AGSF)	0.8278	0.8555	0.8656	0.7996	0.8272	0.8338
w/o (TAAP + PACL)	0.7668	0.7838	0.7893	0.7696	0.7841	0.7902
w/o (AGSF)	0.8377	0.8584	0.8709	0.8157	0.8313	0.846
w/o (PACL)	0.9013	0.945	0.9639	0.906	0.9517	0.9676
PromptHash	**0.9313**	**0.9759**	**0.9931**	**0.9381**	**0.9761**	**0.9934**

Research

Experiments on the NUS-WIDE dataset show that the TAAP module alleviates text semantic truncation and enhances text feature representation. The AGSF module adaptively fuses multimodal semantics, retaining useful information and filtering out redundancy. Aligning global and local prompt tokens (PACL) further highlights relevant semantics and suppresses background noise. Overall, PromptHash demonstrates superior performance and robustness.

PR Curve



Precision-Recall (PR) Curve Results on the Three Benchmark Datasets

Conclusion & Future Work

优化文本

去噪融合

前沿技术

In our study, we observed that all three commonly used public cross-modal hashing datasets contain substantial noise that adversely affects retrieval performance, with the issue being particularly prominent in the MS COCO dataset. Moreover, the text annotations in the other two datasets are based on discrete word labels rather than natural sentences. For future work, we propose to optimize the original annotated texts and images by leveraging diffusion models for image enhancement and employing large language models such as GPT-4 to reconstruct text annotations into prompt-like sentences. Additionally, we will explore integrating weakly-supervised segmentation techniques, introducing CAM-based image prompts as retrieval targets, to further improve retrieval performance.

2025

Thank you for watching