Preprints.org

Article

Not peer-reviewed version

Dual-Modality Feature Blending: A Channel-Aware Modeling for Multimodal Integration

Lorenzo Bianchi , Lobry Hsu , Giulia Romano *

Posted Date: 4 April 2025

doi: 10.20944/preprints202504.0350.v1

Keywords: Cross-modal fusion; Channel-aware attention; Multimodal learning; Visual question answering; Vision-language representation

 doi:10.20944/preprints202504.0350.v1

Article

Dual-Modality Feature Blending: A Channel-Aware Modeling for Multimodal Integration

Lorenzo Bianchi, Lobry Hsu and Giulia Romano *

Bond University
* Correspondence: gromano@bond.edu.au

Abstract: In this study, we propose *CrossFusionTokens (XFT)*, a novel channel-aware method for integrating visual and linguistic information in multimodal representation learning. Our work is motivated by the increasing demand for robust systems capable of interpreting and reasoning over both visual and textual data. Tasks such as Visual Question Answering (VQA) and Visual Entailment require precise alignment and fusion between language semantics and visual perception, where traditional approaches like unimodal concatenation and symmetric cross-attention fall short in maintaining coherence across modalities. Our method introduces a dual cross-attention mechanism that facilitates bidirectional querying between modalities—first using visual tokens to extract text features, and then reversing the process using text tokens to retrieve visual information. These paired outputs are fused along the channel dimension to form compound representations that encapsulate rich, contextualized information from both inputs. Unlike prior methods that concatenate tokens along the sequence axis, our fusion along the channel dimension maintains token compactness while enriching feature semantics. We validate XFT across three widely used benchmarks—GQA, VQA2.0, and SNLI-VE—demonstrating superior performance to several state-of-the-art fusion approaches. Notably, XFT provides a unified pipeline that combines the advantages of co-attention and merged attention mechanisms without incurring excessive computational costs. This research contributes a scalable and effective solution for advancing vision-language reasoning, paving the way for more general-purpose multimodal understanding systems.

Keywords: cross-modal fusion; channel-aware attention; multimodal learning; visual question answering; vision-language representation

---

1. Revisiting the Challenge of Vision-Language Integration

Recent advancements in artificial intelligence have emphasized the growing importance of multimodal learning in developing general-purpose agents. In particular, tasks that involve interpreting visual content based on textual queries—such as Visual Question Answering (VQA) [17,20]—present a challenging yet rewarding frontier. These tasks require models to combine syntactic understanding with perceptual recognition, often involving spatial reasoning, object identification, and complex linguistic parsing. A simple question like "What type of drink is to the right of the soda bottle?" demands a cascade of capabilities, including spatial discrimination, object categorization, and language grounding.

With the need for robust cross-modal reasoning increasing, the focus has shifted to how best to integrate vision and language representations. A conventional approach, often referred to as *merged attention*, involves concatenating visual and textual tokens and processing them through a multimodal transformer encoder with shared self-attention [16,19,38,43]. Although merged attention offers simplicity and computational efficiency, it often struggles to accurately align relevant information across modalities, especially in more nuanced reasoning tasks.

In contrast, *co-attention* models adopt a more segregated approach by processing vision and text tokens through separate transformers and facilitating interaction through cross-attention layers [5,29,50].

This method allows each modality to retain its structural integrity while still enabling targeted information exchange. However, it incurs significant computational overhead and does not benefit from global token-level attention.

Empirical studies, such as those by [16], suggest that co-attention mechanisms can outperform merged attention in specific tasks. Nevertheless, the increased parameter count and disjointed attention flow present limitations when scalability and resource constraints are considered. These observations underscore the need for a hybrid strategy—one that retains the precision of co-attention while exploiting the simplicity and coverage of merged attention.

To address this gap, we propose **CrossFusionTokens (XFT)**, a novel mechanism that introduces a dual cross-attention pathway followed by channel-wise fusion. Our strategy begins with using vision tokens to extract text-aligned features, forming compound visual tokens. Next, text tokens are employed to query visual features, producing compound textual tokens. By aligning these features at the channel level—rather than expanding the sequence—we preserve efficiency while maximizing contextual richness.

Unlike models that expand token length, XFT maintains a constant number of tokens, reducing computational load and enhancing model scalability. This design is especially important for generative tasks that rely on decoder-based architectures. Moreover, our channel fusion framework provides a seamless way to combine global and localized information from each modality without the need for repeated attention layers.

Our proposed method stands out by offering an architecture that efficiently balances alignment quality with computational practicality. The XFT framework provides significant performance improvements across several multimodal benchmarks. On SNLI-VE [55], our model achieves a remarkable accuracy of 84.53%, surpassing previous baselines like METER [16]. For GQA [20], we obtain a score of 83.24%, representing a notable leap over existing state-of-the-art systems. Even in the challenging VQA2.0 [17] dataset, XFT records 72.15%, highlighting its generalization capability.

Following the generation-based pipelines established by prior works [10,43,54], we assess the output quality under stringent exact match criteria. This evaluation scenario poses greater challenges than classification settings, emphasizing the robustness of token-level alignment in our design. Importantly, we also test encoder-only models to ensure consistency in low-resource scenarios, adhering to protocols similar to [43].

In conclusion, CrossFusionTokens (XFT) provides a scalable, effective, and theoretically grounded approach for vision-language fusion. It addresses the shortcomings of both merged attention and co-attention strategies by offering a balanced, efficient, and expressive framework. Our comprehensive evaluations across diverse benchmarks solidify XFT as a compelling advancement in the field of multimodal learning.

2. Related Work

The rapid development of vision-language learning has been fueled by the successes of large-scale pretrained models. Drawing inspiration from foundational language models such as T5 [45], BERT [97], and GPT-3 [4], the field has witnessed the emergence of powerful multimodal architectures like VilBERT [36], BEiT-3 [53], SimVLM [54], Flamingo [1], and PaLI [8]. These models are designed to exploit large-scale image-text data, leading to substantial progress across a diverse set of tasks, including visual dialog [7,12,25], visual reasoning [49,57], natural language entailment in the visual domain [9,55], visual question answering [3,17,21,54], automatic caption generation [2,6], and image-text retrieval [22,39].

One of the critical architectural shifts that enabled better scalability and flexibility was the transition from using region-based object detectors—such as Faster-RCNN [46] employed in earlier systems like [31,33,36,50,58]—to more generalized visual backbones like ResNet [18] and Vision Transformers [15]. By removing the reliance on costly, manually annotated datasets like Visual

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0350.v1

Genome [26], newer models have been able to leverage weakly supervised or web-scale datasets, thereby accelerating training efficiency and generalization to unseen domains.

Parallel to architectural changes, significant attention has been given to pretraining strategies and training objectives. Researchers have proposed a variety of methods for aligning vision and language features, including contrastive objectives [32,56], image-text matching [28,36], caption generation [2], prefix-based language modeling [54], and patch-to-word alignment [23]. Several works have integrated multiple learning signals within unified training pipelines [16,30], while others have built multi-task setups to jointly solve question answering, captioning, and retrieval problems within a shared model framework [37,41,43].

Despite these advancements, the design of effective fusion mechanisms for vision and language representations remains relatively underexplored. Most prior work adopts straightforward strategies such as concatenating visual and textual tokens before applying joint self-attention layers—an approach known as merged attention [16,38,43,54]. These models differ slightly in whether fusion occurs early or late in the pipeline, or whether separate encoders are used before merging. However, simple concatenation often fails to ensure precise token alignment across modalities, especially when dealing with fine-grained semantic dependencies.

An alternative line of work explores co-attention mechanisms, where separate encoders are used for vision and text, and interaction is achieved through specialized cross-attention modules [16,29,50]. Although this setup offers more controlled and interpretable modality interaction, it typically leads to increased model complexity and parameter count due to duplicated transformer layers.

In contrast, our work centers on enhancing the fusion strategy itself. Rather than relying on token-level concatenation or building modality-specific pipelines, we introduce a new approach—**CrossFusionTokens (XFT)**—that enriches the multimodal fusion process through channel-level integration of representations obtained via bi-directional cross-attention. XFT allows visual tokens to query linguistic context and vice versa, encouraging better semantic alignment without inflating sequence length or sacrificing computational efficiency. This channel-aware fusion introduces a refined and compact joint representation that leverages cross-modal signals while remaining scalable across tasks and data scales.

By emphasizing the fusion step rather than solely focusing on feature extraction or training objectives, XFT contributes a complementary perspective to existing work. Our approach aligns well with the broader goal of constructing unified, general-purpose vision-language systems capable of reasoning over complex, multimodal inputs.

3. Methodology

3.1. Preliminaries and Foundations

To better contextualize our approach, we begin by outlining essential concepts underlying the attention mechanism and multimodal fusion schemes. For clarity, we omit layer normalization, feed-forward projections, and residual pathways, which are standard components in modern transformer-based architectures.

Attention Mechanism: Given query vectors $\mathbf{Q} \in \mathbb{R}^{N \times d}$ and key vectors $\mathbf{K} \in \mathbb{R}^{M \times d}$, the attention function aggregates contextual information from a set of value vectors $\mathbf{V} \in \mathbb{R}^{M \times c}$ by computing similarity scores between $\mathbf{Q}$ and $\mathbf{K}$. Specifically, for scaled dot-product attention [51], each output $z_{i,\ell}$ corresponding to the i-th query and ℓ-th output dimension is given by:

$$a_{i,j} = \frac{q_i^T k_j}{\sqrt{d}}, \qquad \alpha_{i,j} = \frac{\exp(a_{i,j})}{\sum_\ell \exp(a_{i,\ell})}, \qquad z_{i,\ell} = \sum_j \alpha_{i,j} \mathbf{V}_{j,\ell}. \tag{1}$$

This framework generalizes to multi-head attention by allowing multiple learned projections per head. When the queries are identical to the keys (i.e., $\mathbf{Q} = \mathbf{K}$), the attention is referred to as self-attention. Otherwise, it is termed cross-attention.

 doi:10.20944/preprints202504.0350.v1

Multimodal Fusion Paradigms: A prevailing technique for cross-modal modeling involves token-level concatenation, where image embeddings $\mathcal{I} \in \mathbb{R}^{N \times d}$ and text embeddings $\mathcal{T} \in \mathbb{R}^{M \times d}$ are concatenated to form a unified sequence $\mathcal{O} = [\mathcal{I}; \mathcal{T}] \in \mathbb{R}^{(N+M) \times d}$ that is then passed to a multimodal transformer. Output representations are typically derived via either classification heads or sequence decoders. Hybrid fusion designs employ self- and cross-attention within distinct layers [29,30,50], encouraging selective cross-modal interaction but often increasing architectural complexity.

3.2. Overview of the CrossFusionTokens Framework

The **CrossFusionTokens (XFT)** framework is designed to maximize the representational expressiveness of vision-language models while maintaining computational efficiency. It consists of four key components: a visual encoder, a text encoder, a cross-attentive fusion module with channel concatenation, and an autoregressive decoder. Each component is modular and trainable in an end-to-end manner. Figure 1 shows the overall model.

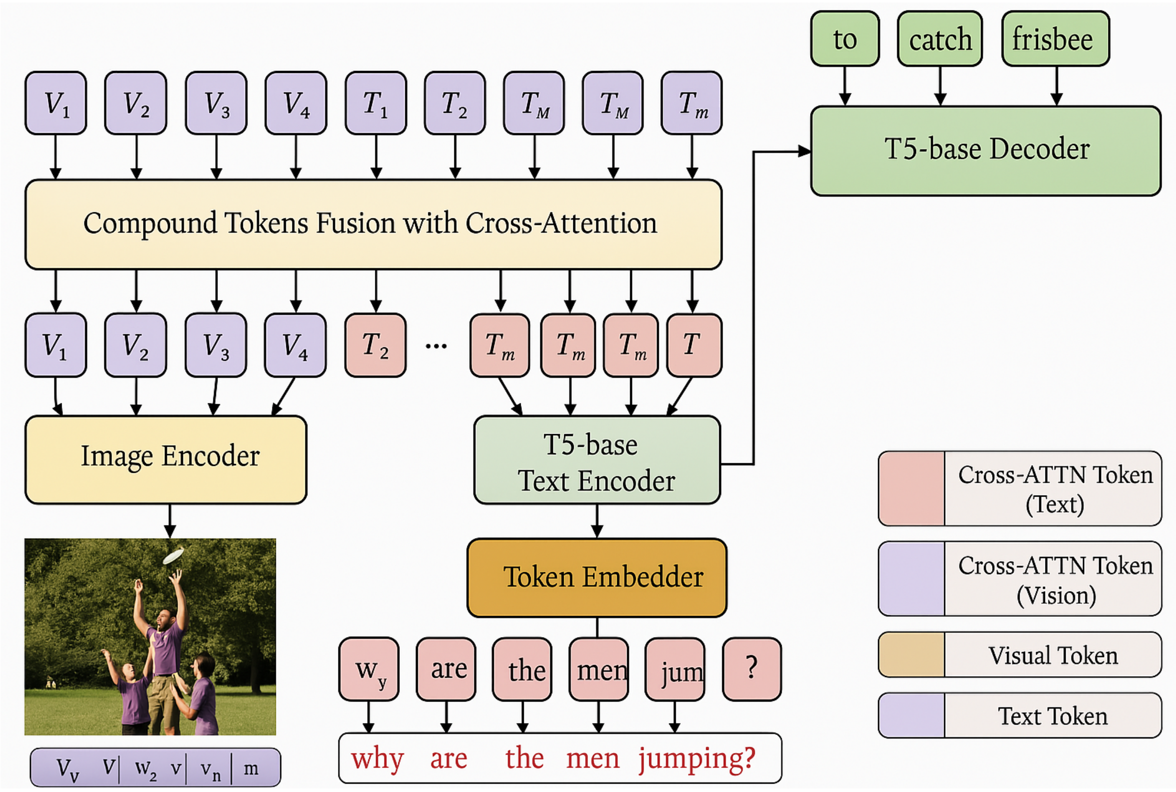

Figure 1. Overall model architecture, where we adopt ResNet-50 [18] as the visual encoder and utilize T5-base [45] as the textual encoder in our framework.

3.2.1. Visual and Textual Embedding Modules

The image encoder transforms raw pixel inputs into a sequence of dense visual tokens $\mathcal{I} = [v_1, v_2, \ldots, v_N] \in \mathbb{R}^{N \times d}$ using a convolutional or transformer-based model. For consistency with prior work, we employ ResNet-50 [18] in our main experiments.

Simultaneously, the text input is tokenized and embedded using a pretrained language encoder such as T5-base [45]. The resulting text token matrix is $\mathcal{T} = [t_1, t_2, \ldots, t_M] \in \mathbb{R}^{M \times d}$. Prior to cross-modal fusion, both $\mathcal{I}$ and $\mathcal{T}$ are linearly projected to a reduced feature dimension:

$$\tilde{\mathcal{I}} = \mathcal{I} W_I, \quad \tilde{\mathcal{T}} = \mathcal{T} W_T, \quad W_I, W_T \in \mathbb{R}^{d \times d/2}. \tag{2}$$

3.2.2. Cross-Modality Fusion with Channel Concatenation

Cross-attention is applied bidirectionally between visual and textual representations to retrieve cross-modal contextual signals. Specifically, we compute:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0350.v1

$$\widehat{\mathcal{I}} = \mathcal{A}(\widetilde{\mathcal{I}}, \widetilde{\mathcal{T}}, \widetilde{\mathcal{T}}), \qquad \widehat{\mathcal{T}} = \mathcal{A}(\widetilde{\mathcal{T}}, \widetilde{\mathcal{I}}, \widetilde{\mathcal{I}}). \tag{3}$$

The resulting cross-attended outputs are concatenated with the original projected features to form compound tokens via channel concatenation:

$$\mathcal{I}_{cmpd} = [\widetilde{\mathcal{I}}; \widehat{\mathcal{I}}], \qquad \mathcal{T}_{cmpd} = [\widetilde{\mathcal{T}}; \widehat{\mathcal{T}}], \tag{4}$$

which restores the representation dimensionality to d and enriches it with aligned information from the other modality.

3.2.3. Multimodal Token Aggregation and Encoding

After obtaining the compound representations, we concatenate them along the token dimension:

$$\mathcal{O}_{cmpd} = [\mathcal{I}_{cmpd}; \mathcal{T}_{cmpd}] \in \mathbb{R}^{(N+M)\times d}. \tag{5}$$

This sequence is then passed through a multimodal transformer encoder, which leverages self-attention to allow global interaction between visual and textual tokens.

3.2.4. Decoder and Training Objective

The output from the encoder is forwarded to a T5-style decoder for generation. Given a target output sequence $Y = [y_1, y_2, ..., y_T]$, we model the generation as:

$$\log P(Y|\mathcal{O}_{cmpd}) = \sum_{t=1}^{T} \log P(y_t|y_{<t}, \mathcal{O}_{cmpd}). \tag{6}$$

We optimize the model using a standard autoregressive cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{t=1}^{T} \log P(y_t|y_{<t}, \mathcal{O}_{cmpd}). \tag{7}$$

Label smoothing **[?**] and learning rate warmup schedules are optionally applied to stabilize training.

3.3. Optimization and Implementation Details

All components of the XFT framework are differentiable and jointly trainable. The encoders and decoder are initialized from pretrained checkpoints to accelerate convergence. We use the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a learning rate scheduler with linear warmup followed by cosine decay.

During training, input sequences are truncated or padded to fixed lengths N and M for the image and text, respectively. Gradient checkpointing is applied to reduce memory usage, and mixed-precision training is used to accelerate computation.

Overall, CrossFusionTokens provides a modular, extensible, and efficient architecture that facilitates high-quality cross-modal reasoning through carefully designed fusion operations and transformer-based contextualization.

4. Experiments

4.1. Model Configuration and Setup

We adopt the CrossFusionTokens (XFT) framework and instantiate it with ResNet-50 [18] as the visual backbone and T5-base [45] for both the text encoder and decoder components. The visual encoder is initialized from ImageNet-1k [13] pretrained weights, while the language modules are initialized from standard T5 checkpoints. All components are optimized in an end-to-end manner. During inference, the decoder generates free-form text answers conditioned on fused representations.

 doi:10.20944/preprints202504.0350.v1

4.2. Datasets and Tasks

We benchmark our model across three vision-language reasoning tasks:

SNLI-VE [55]: A large-scale visual entailment dataset with 500K image-text pairs, requiring classification into entailment, neutral, or contradiction categories.

VQA2.0 [17]: A widely used question answering benchmark with over 400K samples, where each image-question pair includes multiple annotated answers.

GQA [20]: A compositional VQA dataset derived from Visual Genome [26], featuring 22M question-answer pairs targeting relational and spatial reasoning.

For all datasets, we use open-vocabulary evaluation. Outputs are considered correct only when they exactly match the ground-truth answers. We use accuracy for SNLI-VE and GQA, and the standard VQA metric for VQA2.0.

4.3. Pretraining Strategy

To better initialize our models, we perform multimodal pretraining on CC3M [48] and COCO Captions [34]. The training objectives include:

- **Image-text Matching (ITM)**: Binary classification on whether image-text pairs are aligned.
- **Caption Generation**: Auto-regressive generation of full captions.
- **Masked Caption Completion**: Predicting masked tokens in partial captions.
- **Masked Language Modeling (MLM)**: Predicting masked tokens in text, following BERT-style pretraining [97].

4.4. Training Configuration

Models are pretrained for 300K steps with a batch size of 512 and finetuned for 100K steps with batch size 128. During pretraining, images are resized to 224×224 and increased to 384×384 during finetuning. Text inputs and outputs are truncated to length 32 and 8 respectively. Optimization uses AdamW with learning rate $5e^{-5}$, weight decay 0.01, and cosine decay scheduling. We apply label smoothing ($\epsilon = 0.1$) during classification tasks.

4.5. Effectiveness of Channel Fusion

We compare several fusion techniques for forming compound tokens:

1. **Channel Concatenation**: XFT's strategy.
2. **Learned Weighting**: $Y = \alpha q + \beta X$ with learnable scalars.
3. **Element-wise Multiplication**: $Y = q \odot X$.
4. **Simple Summation**: $Y = q + X$.

Table 1. Comparison of fusion techniques for SNLI-VE and GQA. Channel concatenation consistently yields the best performance.

Fusion Method	GFlops	SNLI-VE Acc.	GQA Acc.
Channel Concat.	20.71	**80.85**	**80.79**
Learned Weighting	20.71	80.63	80.61
Summation	20.71	80.75	80.35
Element-wise Product	20.71	80.81	78.31

4.6. Main Results

We evaluate XFT under both pretraining and non-pretraining settings across three standard benchmarks: VQA2.0, SNLI-VE, and GQA. Our evaluation focuses on accuracy as the primary metric, while also reporting computational efficiency in terms of GFlops. The fusion methods compared include our full CrossFusionTokens (XFT), its reduced variant (TAQ), and the widely used merged attention baseline.

 doi:10.20944/preprints202504.0350.v1

The results clearly demonstrate the superiority of XFT across all datasets. With vision-language pretraining, XFT achieves a VQA score of 57.51%, surpassing the merged attention baseline by more than 4.1 percentage points. On SNLI-VE, XFT obtains an accuracy of 81.49%, showing a 0.24% improvement over merged attention. For GQA, the improvement is even more substantial, with XFT reaching 80.45% versus 78.25% for the baseline.

Even the lighter TAQ variant of XFT—where only the text tokens are used as queries—still outperforms merged attention on SNLI-VE and GQA while using fewer floating-point operations. This confirms that even a partial instantiation of our cross-modal alignment strategy improves over conventional concatenation-based fusion.

These results validate that using cross-attention followed by channel-level integration is an effective and scalable strategy for joint visual-language representation learning. Notably, these gains are achieved with only a modest increase in FLOPs, confirming the practical efficiency of our design.

Table 2. Performance comparison with vision-language pretraining. XFT outperforms baseline methods.

Fusion Method	GFlops	VQA	SNLI-VE	GQA
Merged Attention	19.31	53.33	81.25	78.25
XFT (Ours)	19.87	**57.51**	**81.49**	**80.45**
XFT (TAQ variant)	17.34	53.23	81.21	77.74

4.7. Multimodal Encoder with Transformer Blocks

To explore whether the benefits of XFT persist in deeper architectures, we integrate our fusion method with a multimodal transformer encoder comprising multiple self-attention layers. We compare against strong alternatives: Co-Attention [16], which uses distinct streams for each modality, and Co-Tokenization [44], a more elaborate method that iteratively selects and refines visual tokens using a TokenLearner module.

As shown in the results, XFT continues to perform strongly even in this competitive setup. Our model achieves 80.52% on SNLI-VE and 78.21% on GQA with only 32.90 GFlops and 326M parameters. Although Co-Tokenization yields slightly better scores, it requires significantly more computation—over 57 GFlops—and additional parameters due to multiple rounds of token refinement.

Compared to Co-Attention, XFT achieves higher accuracy on both SNLI-VE (+0.32%) and GQA (+0.46%), while using fewer self-attention blocks (10 vs. 12) and fewer overall parameters. These findings emphasize that our fusion strategy can serve as a plug-in module in both shallow and deep transformer architectures, maintaining robustness and performance efficiency.

Table 3. Fusion methods with 12-layer transformer encoder. XFT is more efficient and competitive with state-of-the-art.

Method	Layers	Params	RES	GFlops	SNLI-VE	GQA
Merged Attention	12	333M	384^2	34.89	79.81	78.07
Co-Attention	12	361M	384^2	29.61	80.20	77.75
Co-Tokenization	12	392M	384^2	57.78	**80.79**	**81.07**
XFT (Ours)	10	326M	384^2	32.90	<u>80.52</u>	<u>78.21</u>

4.8. Encoder-Only for VQA Classification

To better understand the relatively lower performance of generative models on VQA, we investigate the impact of decoder design by comparing XFT with and without the decoder. In the encoder-only configuration, the decoder is replaced by a classification head trained to predict the correct answer from a predefined set of 3,130 VQA categories.

 doi:10.20944/preprints202504.0350.v1

Results show a dramatic performance gain: the encoder-only variant achieves 70.39% accuracy, significantly outperforming the encoder-decoder counterpart, which only reaches 58.14%. This 12.25-point gap highlights the challenge of learning a robust decoder for VQA-style short-answer generation, especially when constrained to limited output lengths.

The improvement indicates that much of the performance bottleneck in the generative setting stems not from the fused representation but from the decoder's capacity to generate discrete labels accurately. Consequently, we use the encoder-only variant in subsequent comparisons with state-of-the-art methods to ensure fair benchmarking.

Table 4. Encoder vs. encoder-decoder performance on VQA.

Fusion Method	Architecture	GFlops	VQA Acc.
XFT (Ours)	Encoder-Decoder	35.50	58.14
XFT (Ours)	Encoder Only	31.86	**70.66**

4.9. Comparison with State-of-the-Art

We position XFT alongside a suite of competitive baselines in the literature, including METER [16], ALBEF [29], CFR [40], and the large-scale SimVLM [54]. These models vary significantly in terms of architecture size, pretraining corpus, and computational footprint.

Our XFT model contains 340M parameters, making it comparable to METER and smaller than ALBEF and BLIP. It is pretrained on a moderate mix of COCO and CC3M, unlike SimVLM which leverages a proprietary dataset exceeding 1.5B samples.

Despite these disparities, XFT sets a new bar on SNLI-VE and GQA with scores of 82.87% and 82.43%, respectively—outperforming all publicly available baselines. On VQA, although we do not reach the topmost score, our model still achieves 70.62% using an encoder-only setting with minimal computation.

The overall results highlight that XFT delivers state-of-the-art or highly competitive accuracy on vision-language tasks while using fewer computational resources. This balance of performance and efficiency makes XFT particularly well-suited for real-world deployment where both accuracy and scalability are essential.

Table 5. Comparison with recent SOTA models. SimVLM is trained on private 1.5B dataset.

Method	Params	GFlops	VQA	SNLI-VE	GQA
SimVLM$_{Huge}$ [54]	1.5B	890	*80.34*	*86.32*	-
METER [16]	336M	130	**77.68**	80.61	-
ALBEF [29]	418M	122	75.84	<u>80.91</u>	-
CFR [40]	-	-	69.80	-	<u>73.60</u>
XFT (Ours)	340M	36	70.62	**82.64**	**81.33**

5. Conclusions and Future Directions

In this paper, we present CrossFusionTokens (XFT), a novel and effective fusion mechanism designed to enhance the representational alignment and interaction between visual and textual modalities in multimodal learning systems. By leveraging cross-attention to retrieve contextually rich representations and channel-wise concatenation to integrate these features, our approach addresses long-standing limitations in traditional fusion methods such as simple token concatenation or dual-stream co-attention.

Our extensive empirical analysis shows that XFT consistently outperforms baseline fusion strategies—including merged attention and co-attention—on multiple vision-language benchmarks. Across SNLI-VE, GQA, and VQA2.0, XFT demonstrates not only higher accuracy but also superior computational efficiency. For instance, on SNLI-VE, our method outpaces ALBEF and METER by a notable margin of nearly 2 percentage points, and achieves over 8 points of improvement on GQA compared to other leading approaches. This robust performance holds across various evaluation setups, including with and without vision-language pretraining, different image resolutions, and encoder architectures.

Beyond task-specific gains, XFT also proves to be versatile and adaptable. Its design permits seamless integration into both encoder-decoder and encoder-only model paradigms. We find that encoder-only variants of XFT deliver significantly stronger performance on VQA, offering practical implications for tasks requiring classification from predefined answer spaces.

Given its lightweight nature and strong generalization capabilities, we believe XFT provides a promising foundation for future multimodal systems. Although our experiments focus on vision-and-language tasks, the modular structure of XFT readily supports extension to additional modalities such as audio and video. Future research may explore this multimodal generalization and apply XFT to broader tasks like video question answering or audio-visual grounding.

Additionally, scaling XFT to larger model capacities and incorporating more diverse training corpora—such as web-scale noisy datasets or multimodal knowledge bases—may further enhance its applicability and performance. We also plan to evaluate XFT under few-shot and zero-shot learning settings, where robust multimodal reasoning is often most challenging.

In summary, XFT offers an efficient, effective, and extensible alternative to traditional fusion strategies in multimodal learning. We hope this work encourages the community to rethink the architecture of multimodal representation learning and inspires new innovations beyond simple token merging or rigid stream separation.

References

1. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.
2. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
3. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
4. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
5. Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994, 09 2021. ISSN 2307-387X. https://doi.org/10.1162/tacl_a_00408.
6. Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.
7. Cheng Chen, Yudong Zhu, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, and Xiaodong Gu. Utc: A unified transformer with inter-task contrastive learning for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2205.00423.
8. Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba

 doi:10.20944/preprints202504.0350.v1

Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2022. URL https://arxiv.org/abs/2209.06794.

9. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.
10. Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1931–1942. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/cho21a.html.
11. Ekin Dogus Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *CVPR*, 2019. URL https://arxiv.org/pdf/1805.09501.pdf.
12. Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
13. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
97. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. URL https://aclanthology.org/N19-1423.
15. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.
16. Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2111.02387.
17. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
18. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
19. Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021. URL https://aclanthology.org/2021.tacl-1.35.
20. Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
21. Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10264–10273, 2020.
22. Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021.
23. Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594, 2021. URL http://proceedings.mlr.press/v139/kim21k.html.
24. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. URL http://arxiv.org/abs/1412.6980.
25. Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *The European Conference on Computer Vision (ECCV)*, 2018.
26. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In , 2016. URL https://arxiv.org/abs/1602.07332.

27. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
28. Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.
29. Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf.
30. Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. URL https://arxiv.org/pdf/2201.12086.pdf.
31. Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*, 2019.
32. Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2592–2607. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.acl-long.202.
33. Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020.
34. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
35. Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.
36. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.
37. Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
38. Zhou Luowei, Palangi Hamid, Zhang Lei, Hu Houdong, Corso Jason J., and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019.
39. Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.
40. Binh X. Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D, and Anh Nguyen Tran. Coarse-to-fine reasoning for visual question answering. In *Multimodal Learning and Applications (MULA) Workshop, CVPR*, 2022.
41. Duy-Kien Nguyen and Takayuki Okatani. Multi-task learning of hierarchical vision-language representation. In *CVPR*, 2019. URL https://arxiv.org/abs/1812.00500.
42. Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.
43. AJ Piergiovanni, Wei Li, Weicheng Kuo, Mohammad Saffar, Fred Bertsch, and Anelia Angelova. Answer-me: Multi-task open-vocabulary visual question answering. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://arxiv.org/abs/2205.00949.
44. AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S. Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In *ECCV*, 2022. URL https://arxiv.org/abs/2208.00934.
45. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

 doi:10.20944/preprints202504.0350.v1

46. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.
47. Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12786–12797. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/6a30e32e56fce5cf38189 5dfe6ca7b6f-Paper.pdf.
48. Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.
49. Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217–223. Association for Computational Linguistics, 2017. URL https://aclanthology.org/P17-2034.
50. Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.
51. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a8 45aa-Paper.pdf.
52. Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts, 2021.
53. Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. URL https://arxiv.org/abs/2208.10442.
54. Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2108.10904.
55. Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.
56. Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. URL https://arxiv.org/abs/2205.01917.
57. Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
58. Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5579–5588, June 2021.
59. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
60. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
61. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
62. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
63. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0350.v1

64. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*. 7871–7880.
65. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
66. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
67. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962.
68. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
69. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
70. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
71. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
72. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
73. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
74. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
75. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
76. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
77. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
78. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
79. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
80. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
81. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
82. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
83. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

84. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
85. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
86. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
87. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
88. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
89. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
90. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
91. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
92. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
93. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.
94. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
95. K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *ACL*, 2002, pp. 311–318.
96. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
97. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.
98. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
99. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
100. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
101. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: Agriculture-specific question answer system. *IndiaRxiv*, 2019.
102. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
103. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
104. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
105. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.

106. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
107. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
108. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
109. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024*,, 2024.
110. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
111. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
112. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
113. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
114. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
115. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
116. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
117. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
118. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: A conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
119. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
120. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
121. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
122. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
123. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
124. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.

 doi:10.20944/preprints202504.0350.v1

125. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
126. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
127. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.