

Style-aware contrastive adapters for multi-style image captioning

Riyan Mahmudin, Shintami Chusnul Hidayati*, Nanik Suciati

Department of Informatics, Faculty of Electrical Technology and Intelligent Informatics, Institut Teknologi Sepuluh Nopember, Indonesia

Abstract

InstructBLIP demonstrates a strong ability to follow instructions through its instruction-aware Q-Former model. However, for image captioning tasks, this model is trained with instruction templates that are stylistically homogeneous and have similar descriptive intentions. In multi-style food captioning, different styles demand distinct interpretational emphases on identical visual features: ingredient-focused captions require selective attention to texture and compositional patterns, while aesthetic descriptions prioritize color harmony and spatial arrangement. This heterogeneity poses a challenge: whether a single shared Q-Former adaptation mechanism can optimally extract and weight visual features for fundamentally different descriptive objectives. This study proposes enhanced InstructBLIP architecture integrating parallel Style Adapters for style-specific feature transformations and Style-Conditioned Contrastive Learning to regularize adapter outputs. The finding establishes that effective style-specific adaptation requires architectural complementarity: adapters provide parametric capacity for style-specialized transformations, while contrastive learning enforces geometric constraints that prevent representation collapse across styles. These components exhibit mutual dependency: adapters without regularization drift toward homogeneous representations, while contrastive learning without adaptation capacity creates conflicting optimization pressures between discriminative and generative objectives. Experiments validate this complementarity principle, achieving consistent improvements (CIDEr: 181.4 vs. baseline 179.9, +0.8%; SPICE: +1.9%), with substantial gains for structured styles (ingredient: +3.1% CIDEr). Critically, ablation studies confirm this mutual dependency: contrastive loss alone degrades performance (-2.2% CIDEr) while adapters alone yield minimal gains (-1.4% CIDEr), demonstrating that only their combination effectively stabilizes style-specific adaptations in instruction-aware captioning.

This is an open-access article under the [CC BY-SA](#) license.



Keywords:

Contrastive Learning;
Image Captioning;
InstructBLIP;
Parameter-Efficient Fine-Tuning;
Style Adapter;
Visual Instruction Tuning;
Visual Language Models;

Article History:

Received: September 30, 2025

Revised: April 10, 2026

Accepted: April 24, 2026

Published: October 2, 2026

Corresponding Author:

Shintami Chusnul Hidayati
Department of Informatics,
Faculty of Electrical Technology
and Intelligent Informatics,
Institut Teknologi Sepuluh
Nopember, Indonesia
Email: shintami@its.ac.id

INTRODUCTION

In recent years, Artificial Intelligence (AI) research has experienced rapid development, as evidenced by the large number of AI studies in various disciplines, such as health [1], security [2],

and many other fields. Among these developments, one area that has also seen significant progress is image captioning. Image captioning integrates computer vision and natural

language processing to generate coherent natural language descriptions from visual inputs [3, 4].

Early research in image captioning generally used approaches based on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). CNNs are used for visual feature extraction, whereas RNNs convert these features into text [5]. This combined CNN-RNN model approach has become the foundation of the field of image captioning and has been proven to produce adequate image descriptions [6]. Although the CNN-RNN approach is successful in generating descriptions for images, it has limitations in handling images with diverse visual variations [7]. It is constrained by the sequential nature of the RNN, which limits computational efficiency [8].

Subsequently, transformer-based models emerged as a novel approach in image captioning tasks. Transformer models utilize self-attention mechanisms that allow for the simultaneous processing of all tokens, thereby improving the efficiency and precision of caption generation [9]. One notable innovation in transformer-based technology is the Vision Transformer (ViT). Unlike CNN-based methods, ViT partitions images into small patches, which are subsequently converted into tokens and arranged in a manner analogous to the word sequences in text. This approach enables the model to identify visual patterns and inter-regional relationships within images more comprehensively [10].

Following the development of ViT, the Bootstrapping Language-Image Pre-training (BLIP) model [11] has emerged as one of the vision-language models that is well-suited for image captioning tasks. BLIP uses transformer-based encoders and decoders to strengthen the relationship between visual features and the descriptive text. BLIP demonstrates better performance in various image captioning tasks, especially when the generated descriptions must be rich in information within the image. However, full fine-tuning of the ViT in BLIP requires a large dataset and high computational resources to retrain all model components, thus limiting accessibility and efficiency in adapting to new domains or tasks.

To overcome these limitations, further development resulted in BLIP-2 [12], which introduces the Q-Former component as a lightweight bridge between a trained and frozen vision encoder and a frozen Large Language Model (LLM). The Q-Former enables the model to extract relevant visual features and align them with the LLM's language representation space without retraining these large components from scratch. Despite the significant advances demonstrated by

BLIP-2, these models were not designed to follow explicit instructions in generating captions. In this context, Visual Instruction Tuning (VIT) [13] emerged as an approach that enables models to follow specific instructions in generating captions more aligned with user requirements. InstructBLIP [14] implements this approach by incorporating instruction tokens into the Q-Former, thereby enabling the model to perform instruction-aware feature extraction.

Applying VIT to food images is an interesting option because food images have various complex visual aspects, ranging from ingredients and texture to color and presentation. However, currently available food datasets contain only one descriptive caption per image. This poses a problem when applying Visual Instruction Tuning (VIT), which requires diverse instruction-caption pairs to train the model to follow various types of descriptive instructions. A previous study [15] introduced a pipeline to address the limited availability of datasets for VIT. In that study, a new dataset was created using the GPT-4 LLM model. This approach has been shown to produce high-quality datasets for VIT model training, which can serve as a foundation for overcoming the limitations of VIT datasets in the food domain.

The application of VIT using the InstructBLIP model has demonstrated an excellent ability to follow visual instructions through the instruction-aware Q-Former mechanism. However, this model was trained with relatively homogeneous instruction templates in terms of descriptive style. Therefore, the Q-Former's ability to perform selective feature extraction and adaptive feature weighting in a single global transformation function may not be flexible enough to accommodate the representational emphasis required for multi-style captioning.

The challenge of multi-style captioning has been identified in several previous studies. Previous research has explored controllable generation through structured semantic augmentation [16] and energy-based sampling methods [17]. The first approach utilizes semantic templates to guide caption generation toward specific spatio-semantic aspects, while the second employs constraint functions with gradient-based sampling. However, both methods primarily impose constraints at the output level rather than enabling differential visual feature interpretation at the feature extraction stage.

Research on Parameter-Efficient Fine-Tuning (PEFT) has demonstrated that adapter modules can facilitate task-specific adaptation with minimal parameter addition [18, 19, 20].

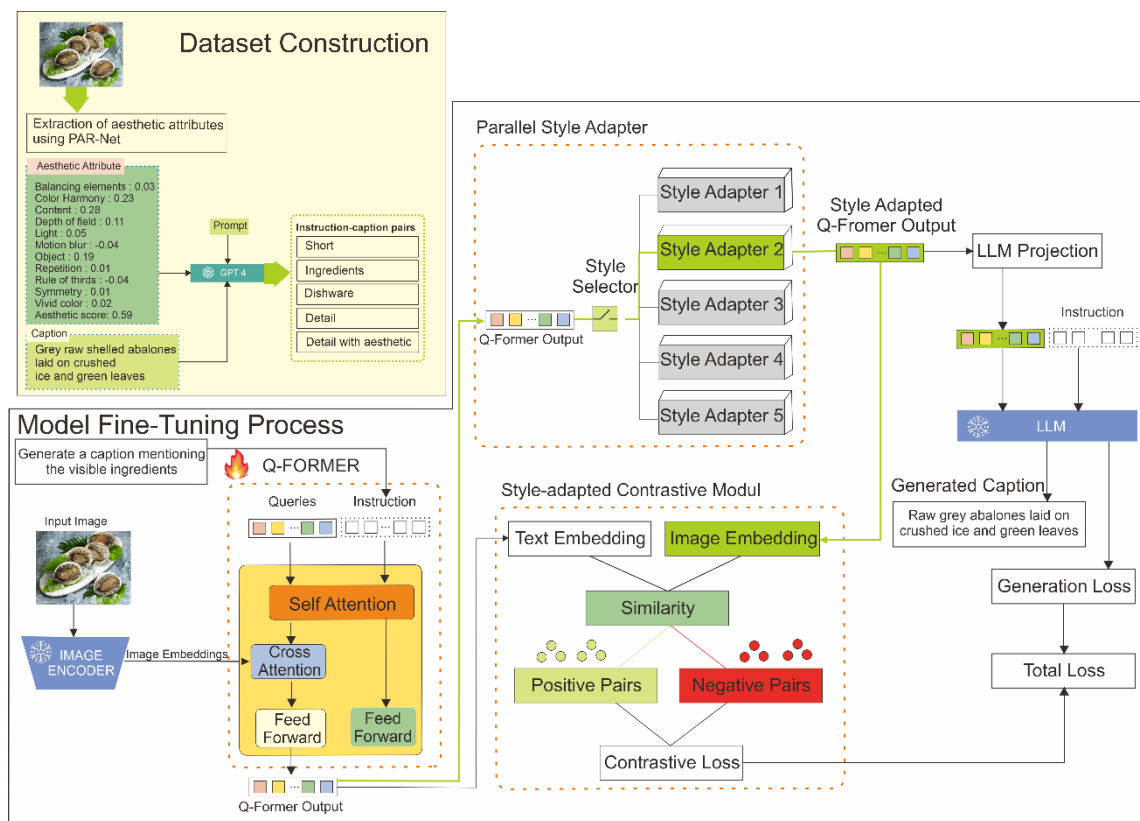


Figure 1. Dataset construction and model fine-tuning process in the modified InstructBLIP model with parallel style adapters and a style-adapted contrastive module

These studies demonstrated that adapters can be employed to adapt pre-trained vision-language models to various novel tasks with competitive results. However, the application of adapters for specific style adjustments in image captioning tasks in the context of VIT has not yet been explored. Previous studies have shown that adapters can capture distinct aspects within a single task. [21] demonstrated that independent adapters can infuse different types of knowledge (such as linguistic and factual) into language models for the same downstream task, with each adapter specializing in a specific knowledge domain. This paradigm aligns with a multi-style captioning scenario, where different caption styles require adapters to specialize in distinct interpretational focuses of identical visual inputs. In addition, contrastive learning has been proven effective in organizing multimodal representation spaces [22, 23]. A study [24] found that contrastive learning that considers style can help models capture different visual content for each style. However, this study focused on style-aware visual encoders in CNN encoder architectures with transformer decoders and has not been explored in the context of VIT with adapters. Nevertheless, the principle of style-conditioned contrastive

learning is relevant to this study's hypothesis that style-conditioned contrastive loss can help prevent adapters from experiencing representation shifts in multi-style captioning adaptation on VIT.

Based on the gaps and findings from previous research, there exists an opportunity to explore the impact of synergistic integration between style-specific adapters and style-conditioned contrastive learning for food captioning tasks.

Unlike prior approaches that apply controllable generation at the output level [16, 17] or adapt vision-language models for single-style tasks [18, 19, 20], this study jointly integrates style-specific parallel adapters and style-conditioned contrastive learning within an instruction-tuned vision-language model for explicit multi-style food image captioning. This study makes two primary contributions:

1. Construction of a multi-style instruction-caption dataset for the food domain that can benefit future research in food image captioning and VIT.
2. Development of an InstructBLIP architecture incorporating style-specific adapters and style-conditioned contrastive learning.

Table 1. Comparison between the Food-500 Cap dataset. Food-500 Cap has only one caption, and the constructed dataset has multi-style captions

Image	Dataset	Instruction	Caption
	Food-500 Cap	-	A green bowl filled with es campur with pink sago and white shaved ice
	Constructed Dataset	Generate a short caption.	A green bowl filled with es campur
		Generate a caption mentioning the visible ingredients.	A green bowl filled with es campur featuring pink sago and white shaved ice
		Generate a food caption describing dishware.	Es campur presented in a vibrant green bowl
		Create a detailed caption that covers the essential elements of the dish.	A green bowl filled with es campur, with pink sago and white shaved ice
	Food-500 Cap	-	A thin slice of pie evenly spread with guacamole and next to a verdant lime.
	Constructed Dataset	Generate a short caption.	A slice of pie with guacamole
		Generate a caption mentioning the visible ingredients.	A thin slice of pie spread with guacamole and accompanied by lime
		Generate a food caption describing dishware.	A slice of pie served on a rustic wooden board next to a vibrant lime
		Create a detailed caption that covers the essential elements of the dish.	A thin slice of pie evenly spread with guacamole and next to a verdant lime.
		Generate a detailed caption that also describes aesthetic qualities.	A thin slice of pie evenly spread with guacamole and next to a verdant lime. The soft natural lighting enhances the fresh colors, while the composition emphasizes the vibrancy of the ingredients, creating an inviting visual appeal

METHOD

This study comprises two primary stages: (1) multi-style dataset construction, and (2) model fine-tuning of a modified InstructBLIP architecture enhanced with parallel style adapters and style-conditioned contrastive learning. The overall methodological framework is illustrated in Figure 1.

Material

Visual Instruction Tuning (VIT) requires a diverse dataset of instruction-caption pairs for each image so that the model can learn to follow different descriptive intentions. However, the currently available food dataset only has one image-caption pair that is not accompanied by instructions. Therefore, this study requires a dataset construction stage. This study includes several stages in dataset construction. The first stage involves selecting food datasets and determining the styles to be applied. This study uses Food-500 Cap [25] as the source dataset, which contains 24,700 food images from 494 categories with human-labeled captions.

Food-500 Cap was selected as the source dataset for several reasons. First, its captions are human-labeled and already contain compositionally rich multi-style information, including ingredient descriptions (blue), food

colors (green), and serving vessel references (orange), as illustrated in Figure 2 of the original paper.

This inherent multi-style composition provides a strong textual foundation for GPT-4-based caption expansion, as the model can derive style attributes from the original caption. Second, the dataset's scale (24,700 images across 494 food categories) ensures sufficient category diversity for multi-style instruction tuning. These characteristics make Food-500 Cap particularly suitable as a source dataset for constructing multi-style instruction-caption pairs in the food domain. However, during the data collection phase, many captions did not have matching images. Therefore, this study considered only categories with at least 40 image-caption pairs to avoid dataset imbalance in each category.



Label: agedashi tofu

Region: Japanese

Caption: Five pieces of golden agedashi tofu with some scallions, carrot strips and wooden fish flowers in a long white plate.

Figure 2. An example from Food-500 Cap

This filtering resulted in 291 categories with 14,490 image-caption pairs.

The multi-style framework in this study encompasses five distinct caption styles, each emphasizing different descriptive aspects: short for brief captions containing the name of the main food, ingredient for captions containing the name of the food accompanied by information on other ingredients in the image, dishware for captions emphasizing the serving container, detail for complete captions containing a combination of all the previous styles, and the last style is detail accompanied by the aesthetic value of the image. The addition of aesthetic value is based on studies of consumer psychology showing that the visual appeal of food advertisements, reflected in a balanced composition, vibrant color harmony, and high-quality lighting, influences perceptions of taste and freshness and even increases the intention to purchase a food item [26]. Therefore, captions that reflect the aesthetic value of food images are important.

To integrate visual aesthetic information, aesthetic attribute scores were extracted using the Photo Aesthetic Ranking Network with Attributes and Content Adaptation (PAR-Net) model. This study used the complete PAR-Net architecture (Reg+Rank+Att+Cont variant), which performed the best in the study. This model predicts 11 aesthetic attributes: balance elements, rule of thirds, symmetry, repetition, depth of field, lighting quality, motion blur, color harmony, vivid colors, objects, and content. Details of the PAR-Net implementation and architecture can be found in [27].

The final stage of dataset construction was the generation of multi-style instruction-caption pairs using GPT-4 via the OpenAI API. The original Food-500 Cap captions and aesthetic attributes extracted in the previous stage were used as inputs for the GPT-4 model. To ensure that the instruction-caption pairs corresponded to the expected multi-style, a prompt was applied to guide the GPT-4 model, which has been documented in EmoViT with adjustments in this study. This stage produces a new dataset consisting of 72,450 pairs, as shown in Table 1, and Table 2 shows a statistical comparison of both datasets.

Table 2. Dataset Statistics Comparison

Aspect	Food-500 Cap	Constructed Dataset
Categories	494	291
Images	24,700	14,490
Caption/Image	1	5 (multi-style)
Total pairs	24,700	72,450
Style diversity	No	5 types

Methods

The fine-tuning process employs InstructBLIP as the base architecture, enhanced with two novel components that constitute the primary contributions of this work: parallel style adapters for style-specific feature transformation and a style-conditioned contrastive learning module for representation regularization. During training, only the Q-Former, style adapters, and contrastive module parameters are optimized. At the same time, the vision encoder and language model remain frozen to preserve their pre-trained knowledge and reduce computational requirements.

Visual Feature Extraction Through Frozen Vision Encoder

The training process begins by encoding input food images into dense visual representations. Each input image $I \in \mathbb{R}^{H \times W \times 3}$ is fed into a frozen EVA-CLIP Vision Transformer (ViT-g/14), which partitions the image into non-overlapping patches. These patches are flattened and linearly projected into high-dimensional embedding vectors that serve as input to the transformer layers.

The visual encoding proceeds through multiple transformer layers, where each layer ℓ processes the output from the previous layer through two sequential operations. First, multi-head self-attention (MSA) is applied with layer normalization and a residual connection. Second, a feed-forward network (MLP) is applied with layer normalization and residual connection, as expressed in (1) and (2):

$$z'_\ell = \text{MSA}(\text{LN}(z_{\ell-1})) + z_{\ell-1} \quad (1)$$

$$z_\ell = \text{MLP}(\text{LN}(z'_\ell)) + z'_\ell \quad (2)$$

where $z_{\ell-1}$ represents the feature representation from the previous layer. The output from the final transformer layer is extracted as the visual representation $V \in \mathbb{R}^{N \times d_v}$, where N is the number of image patches, and d_v is the visual feature dimension. All encoder parameters θ_{enc} remain frozen during training.

Instruction-Aware Visual-Language Alignment via Q-Former

The visual features V from the encoder are aligned with instruction embeddings through the Q-Former module. The Q-Former jointly processes learnable query tokens $Q \in \mathbb{R}^{K \times d_q}$ and tokenized instruction embeddings $T \in \mathbb{R}^{L \times d_q}$, concatenated along the sequence dimension, as formulated in (3):

$$Z_0 = [Q; T] \quad (3)$$

where $[\cdot; \cdot]$ denotes concatenation along the sequence dimension. This concatenated representation $Z_0 \in \mathbb{R}^{(K+L) \times d_q}$ then undergoes processing through multiple transformer layers. At each layer, the model performs three critical operations: (i) self-attention among queries and instruction tokens enables the queries to capture instruction semantics, (ii) cross-attention between the concatenated tokens and visual features V allows extraction of instruction-relevant visual content from the encoded image, and (iii) feed-forward transformation applies non-linear mapping to the attended features. These operations incorporate residual connections and layer normalization to maintain a stable gradient flow during training. After L layers, the query tokens are extracted as shown in (4):

$$Q_{out} = Z_L[:, :K, :] \quad (4)$$

where $Q_{out} \in \mathbb{R}^{K \times d_q}$ represents the instruction-aware visual features; this output encodes both visual content from the image V and semantic guidance from the instruction T , making it sensitive to the type of description required. The Q_{out} serves as input to the subsequent style adaptation stage.

Style-Specific Feature Transformation via Parallel Adapters

Different caption styles require distinct representational emphases ingredient-focused captions demand attention to texture and compositional patterns, aesthetic captions prioritize color harmony and spatial arrangement, while dishware descriptions emphasize object boundaries. To enable these style-specific interpretations, instruction-aware features Q_{out} are transformed through parallel style adapters.

The transformation process employs a bottleneck architecture that projects feature to a lower-dimensional space, applies non-linear transformations, and projects back to the original dimension while preserving the original features through residual connections. Given the instruction-aware features $x = Q_{out}$ from the Q-Former, each sample is routed to its corresponding style-specific adapter based on its style label. The transformation proceeds through four sequential stages, as expressed in (5)–(8):

First, the input features are projected to a low-rank bottleneck dimension:

$$x_{down} = xW_{down}^y + b_{down}^y \quad (5)$$

where $W_{down}^y \in \mathbb{R}^{d_q \times r}$ is the style-specific downsampling matrix, r is the bottleneck dimension (where $r \ll d_q$), and superscript y indicates the style category. This downsampling compresses the features from dimension d_q to r .

Second, non-linear activation and dropout regularization are applied to the compressed features:

$$h = \text{Dropout}(\text{GELU}(x_{down}), p) \quad (6)$$

where GELU provides smooth non-linearity and dropout with probability p prevents overfitting by randomly zeroing elements during training.

Third, the transformed features are projected back to the original dimension:

$$x_{up} = hW_{up}^y + b_{up}^y \quad (7)$$

where $W_{up}^y \in \mathbb{R}^{r \times d_q}$ is the style-specific upsampling matrix that expands features from dimension r back to d_q .

Finally, the transformed features are combined with the original input through a residual connection and normalized:

$$x_{adapted} = \text{LayerNorm}(\text{Dropout}(x_{up}, p) + x) \quad (8)$$

The residual connection ($+x$) in (8) ensures that if adapter learning fails or produces suboptimal transformations, the baseline Q-Former features are preserved, preventing catastrophic performance degradation. During batch processing with style labels $y = \{y_1, \dots, y_B\}$, each sample i is routed to its corresponding adapter A_{y_i} , as formulated in (9):

$$x_{adapted}^{(i)} = A_{y_i}(Q_{out}^{(i)}) \text{ for } i = 1, \dots, B \quad (9)$$

where $y_i \in \{1, \dots, S\}$ indicates the style label for sample i , S is the total number of style categories, and B is the batch size. During training, only the adapter corresponding to each sample's style receives gradients, enabling independent style specialization.

Style-Conditioned Contrastive Learning

The style-adapted features $x_{adapted}$ from different adapters may suffer from representation drift, where outputs for different styles collapse into similar feature distributions despite their distinct semantic objectives. To prevent this collapse, we apply style-conditioned contrastive learning that structures the feature space by enforcing style-aware clustering. The

regularization process operates on two types of features: style-adapted visual features x_{adapted} and the instruction representations extracted from the Q-Former output Z_{out} . Both are projected into a shared alignment space where their similarity can be measured and optimized.

For visual features, the style-adapted query tokens are first aggregated through mean pooling across the K query positions and then projected to the alignment space, as expressed in (10):

$$h_i^{\text{img}} = f_{\text{img}}\left(\frac{1}{K} \sum_{k=1}^K x_{\text{adapted}}^{(i)}[k, :]\right) \quad (10)$$

where $x_{\text{adapted}}^{(i)}[k, :]$ denotes the k -th query token of the sample i after style adaptation, and f_{img} is a projection function that maps the aggregated features to the alignment space. This aggregation reduces the sequence of K tokens into a single vector representation.

Similarly, instruction representations are aggregated from the Q-Former output by extracting tokens at positions $K+1$ to $K+L$ (corresponding to instruction token positions in Z_{out}) and applying mean pooling followed by projection, as shown in (11):

$$h_i^{\text{txt}} = f_{\text{txt}}\left(\frac{1}{L} \sum_{\ell=1}^L Z_{\text{out}}^{(i)}[K+\ell, :]\right) \quad (11)$$

where L is the instruction sequence length and f_{txt} is the instruction projection function. Both projection heads f_{img} and f_{txt} are implemented as two-layer perceptrons that transform their inputs through a hidden layer with ReLU activation, as formulated in (12):

$$f(x) = W_2 \cdot \text{ReLU}(W_1 \cdot x + b_1) + b_2 \quad (12)$$

where $W_1 \in \mathbb{R}^{d_q \times d_{\text{align}}}$ and $W_2 \in \mathbb{R}^{d_{\text{align}} \times d_{\text{align}}}$ are weight matrices, and d_{align} is the alignment space dimension. These projections map both visual and instruction features into a common space where their similarity can be compared. To enable scale-invariant similarity computation, the projected features are L2-normalized according to (13):

$$\tilde{h}_i^{\text{img}} = \frac{h_i^{\text{img}}}{\|h_i^{\text{img}}\|_2}, \tilde{h}_i^{\text{txt}} = \frac{h_i^{\text{txt}}}{\|h_i^{\text{txt}}\|_2} \quad (13)$$

where $\|\cdot\|_2$ denotes the L2 norm. This normalization ensures that the similarity scores depend only on the direction of the feature vectors, not their magnitude, which improves training stability. These normalized features are then used to compute pairwise cosine similarities across all

samples in a batch, forming a similarity matrix using (14):

$$S[i, j] = \frac{(\tilde{h}_i^{\text{img}})^T \tilde{h}_j^{\text{txt}}}{\tau} \quad (14)$$

where $S[i, j]$ represents the similarity between the visual features of sample i and the instruction features of sample j , and τ is a learnable temperature parameter that controls the concentration of the similarity distribution. Higher temperature values produce softer distributions, while lower values create sharper distinctions. The contrastive objective is designed to increase the similarity between samples sharing the same style (positive pairs) while decreasing the similarity with those of differing styles (negative pairs). Given style labels $y = \{y_1, \dots, y_B\}$ for the batch, positive pairs are identified through a binary mask defined in (15):

$$M^{\text{pos}}[i, j] = \begin{cases} 1 & \text{if } y_i = y_j \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

where $M^{\text{pos}}[i, j] = 1$ indicates that samples i and j belong to the same style but are different instances. Self-pairs ($i = j$) are excluded to prevent trivial solutions. For each sample i , the average similarity with all its positive pairs is computed as expressed in (16):

$$s_{\text{pos}}(i) = \frac{\sum_j M^{\text{pos}}[i, j] \cdot \exp(S[i, j])}{\sum_j M^{\text{pos}}[i, j]} \quad (16)$$

This aggregates the exponential similarities of all positive pairs and normalizes by the number of positive pairs. The contrastive loss is then computed as the negative log-likelihood of the positive similarity relative to all similarities in the batch, according to (17):

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \log \frac{s_{\text{pos}}(i)}{\sum_k \exp(S[i, k])} \quad (17)$$

where $\mathcal{V} \subseteq \{1, \dots, B\}$ contains only samples that have at least one positive pair in the batch (to avoid division by zero), and $|\mathcal{V}|$ is the number of such samples. This loss encourages the model to produce image-instruction feature pairs that are similar within each style cluster (high $s_{\text{pos}}(i)$) while being dissimilar across different style clusters (low similarity in the denominator), thereby preventing representation drift in the style adapters and maintaining distinct style-specific features.

Caption Generation Via Large Language Model and Loss Computation

The final stage of the training pipeline transforms the style-adapted visual features into natural language captions through a frozen large language model. This stage encompasses three primary processes: projecting visual features into the language model's embedding space, concatenating them with instruction and caption embeddings, and performing autoregressive caption generation with loss computation.

First, the style-adapted features $x_{\text{adapted}} \in \mathbb{R}^{K \times d_q}$ are linearly projected to match the embedding dimension of the language model, as formulated in (18):

$$x_{\text{llm}} = x_{\text{adapted}} W_{\text{proj}} + b_{\text{proj}} \quad (18)$$

where $W_{\text{proj}} \in \mathbb{R}^{d_q \times d_{\text{LLM}}}$ is the projection weight matrix, $b_{\text{proj}} \in \mathbb{R}^{d_{\text{LLM}}}$ is the bias vector, and d_{LLM} is the hidden dimension of the language model. The output $x_{\text{llm}} \in \mathbb{R}^{K \times d_{\text{LLM}}}$ represents the style-transformed visual information in the language model's representational space, enabling the LLM to process visual features as if they were language tokens.

Second, these projected visual embeddings are concatenated with tokenized instruction embeddings and target caption embeddings to form the complete input sequence for the language model, as expressed in (19):

$$E = [x_{\text{llm}}; E_{\text{instruction}}; E_{\text{caption}}] \quad (19)$$

where $E_{\text{instruction}} \in \mathbb{R}^{L_{\text{in}} \times d_{\text{LLM}}}$ contains the embedded instruction tokens, $E_{\text{caption}} \in \mathbb{R}^{L_{\text{out}} \times d_{\text{LLM}}}$ contains the embedded target caption tokens, and $[\cdot; \cdot]$ denotes concatenation along the sequence dimension. The resulting sequence E contains visual information followed by the instruction and ground-truth caption, enabling the language model to learn the mapping from visual-instruction inputs to appropriate captions.

Third, the language model generates captions through autoregressive token-by-token prediction. The probability of generating a complete caption $Y = \{w_1, w_2, \dots, w_T\}$ given the image I and instruction T is computed as the product of conditional probabilities, as shown in (20):

$$P(Y | I, T) = \prod_{t=1}^T P(w_t | w_{<t}, x_{\text{llm}}, T) \quad (20)$$

where w_t is the token at position t , $w_{<t}$ represents all preceding tokens, and the conditional probability at each step depends on the projected visual features x_{llm} , the instruction T , and

previously generated tokens. All language model parameters θ_{LLM} remain frozen during this process, meaning only the projection weights W_{proj} and earlier components (Q-Former and adapters) receive gradients.

The generation loss is computed using cross-entropy over caption tokens only, excluding visual and instruction tokens from supervision, as formulated in (21):

$$\mathcal{L}_{\text{gen}} = -\frac{1}{|\mathcal{C}|} \sum_{t \in \mathcal{C}} \log P(w_t | w_{<t}, x_{\text{llm}}, T) \quad (21)$$

where $\mathcal{C}$ denotes the set of positions corresponding to caption tokens in the sequence E , and $|\mathcal{C}|$ is the total number of caption tokens. This selective masking ensures that the model learns to generate captions without attempting to predict visual features or instructions, focusing optimization solely on caption generation quality.

Finally, the total training objective combines the caption generation loss with the style-conditioned contrastive loss, as expressed in (22):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda \cdot \mathcal{L}_{\text{contrast}} \quad (22)$$

where λ is a hyperparameter that balances the significance of contrastive regularization against the quality of caption generation. During backpropagation, gradients $\mathcal{L}_{\text{gen}}$ flow backward through the projection layer (W_{proj}), style adapters, and Q-Former, optimizing them for caption generation. Simultaneously, gradients from $\mathcal{L}_{\text{contrast}}$ flow backward through the contrastive projection heads (f_{img} , f_{txt}) and style adapters, enforcing style-specific clustering in the feature space. This joint optimization enables the model to generate high-quality, style-appropriate captions while maintaining distinct representations for different caption styles.

RESULTS AND DISCUSSION

Implementation Details

Training was conducted on the proposed multi-style food caption dataset using an 80:10:10 train, validation, and test split. The model uses a frozen EVA-CLIP ViT-g/14 vision encoder and a frozen Vicuna-7B LLM. Style-aware capability is enabled through style adapters (bottleneck $r = 512$) and a contrastive alignment module ($d_{\text{align}} = 512$), with a contrastive loss weight of $\lambda = 0.05$.

Models were trained for 10 epochs with AdamW (weight decay 10^{-2}) and a linear-warmup cosine scheduler (500 warmup steps). A differential learning rate scheme was applied: 1×10^{-5} for base modules (Q-Former, LLM projection) and 1×10^{-4} for newly introduced

components (adapters, contrastive module). Mixed precision (FP16) yielded an effective batch size of 8 on a single NVIDIA A100 GPU.

Caption generation employed beam search with a beam size of 5, a maximum output length of 300 tokens, a repetition penalty of 1.5, and a length penalty of 1.0.

Overall Performance Comparison

As shown in Table 3, the proposed architecture was evaluated under varying configurations of alignment dimension (d_{align}), adapter bottleneck dimension (r), and contrastive loss weight (λ). The optimal configuration employs $d_{\text{align}} = 512$, $r = 512$, and $\lambda = 0.05$, achieving consistent improvements over the baseline: CIDEr increased by 0.8% (181.4 vs. 179.9), ROUGE-L by 0.7% (45.2 vs. 44.9), and SPICE by 1.9% (31.8 vs. 31.2). These improvements align with recent PEFT research in vision-and-language tasks [18], which demonstrated that adapter-based methods can achieve comparable performance using only 4.18% of parameters. The proposed approach achieves similar parameter efficiency, validating that style-specific adaptation can be accomplished without extensive parameter updates.

The BLEU-4 score improved marginally by 0.6%, while METEOR exhibited a slight decrease of 0.9%. This trade-off has been observed in specialized architectures [23], where semantic understanding improvements sometimes sacrifice n-gram precision, an acceptable compromise for multi-style captioning where diverse linguistic expressions are desirable. Increasing the contrastive weight to $\lambda = 0.1$ degraded

performance (CIDEr: 175.9, SPICE: 30.9), indicating that excessive regularization conflicts with the generative objective. Conversely, smaller dimensions ($d_{\text{align}} = 256$, $r = 256$) resulted in marginal improvements (CIDEr: 179.1), suggesting insufficient capacity for capturing fine-grained style distinctions.

Style-specific Performance Breakdown

Table 4 presents per-style evaluation results. The most significant gains occur in structured caption styles: ingredient-focused captions improve by 3.1% in CIDEr (150.7 vs. 146.2) and 3.2% in SPICE (25.6 vs. 24.8), while dishware descriptions improve by 0.4% in CIDEr (309.5 vs. 308.2) and 0.8% in SPICE (40.0 vs. 39.7). Short captions also benefit, with CIDEr increasing by 0.5% (254.4 vs. 253.2) and SPICE by 1.3% (39.2 vs. 38.7). These substantial gains for ingredient-focused captions align with recent work on style-aware visual learning [24], which found that contrastive objectives help models capture different visual content for each style. Ingredient descriptions demand precise attention to compositional elements and textural patterns, which is precisely the type of selective feature attention that style-conditioned contrastive learning encourages.

Detail-oriented captions show mixed results: CIDEr decreases by 0.6% (97.5 vs. 98.1). However, SPICE improves by 2.0% (25.1 vs. 24.6). This pattern suggests the model prioritizes semantic correctness over n-gram overlap when capacity is constrained, similar to observations in parameter-efficient methods for other complex tasks [20].

Table 3. Overall performance across different hyperparameter configurations

Architecture	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Baseline (InstructBLIP)	17.4	22.1	44.9	179.9	31.2
Proposed ($d=256, r=256, \lambda=0.05$)	17.5	21.9	44.9	179.1	31.5
Proposed ($d=512, r=512, \lambda=0.05$)	17.5	21.9	45.2	181.4	31.8
Proposed ($d=512, r=512, \lambda=0.1$)	17.4	21.8	44.6	175.9	30.9

Table 4. Performance breakdown by instruction style

Style	Architecture	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Short	Baseline	20.6	23.1	47.4	253.2	38.7
	Proposed	21.1	23.3	47.1	254.4	39.2
Ingredient	Baseline	13.3	19.7	41.2	146.2	24.8
	Proposed	13.6	19.9	41.8	150.7	25.6
Dishware	Baseline	32.0	26.2	57.1	308.2	39.7
	Proposed	32.4	26.4	57.3	309.5	40.0
Detail	Baseline	10.4	18.3	36.1	98.1	24.6
	Proposed	10.4	17.8	36.2	97.5	25.1
Detail+aesthetic	Baseline	17.6	23.5	42.9	80.1	28.4
	Proposed	17.4	23.9	43.8	80.7	28.9

The SPICE improvement indicates accurate capture of semantic relationships and object attributes, even if phrasing differs from references. Aesthetic-enhanced detail captions demonstrate modest improvements (METEOR +1.7%, ROUGE-L +2.1%, CIDEr +0.7%, SPICE +1.8%), validating that the contrastive module successfully distinguishes aesthetic-oriented instructions. These results confirm that style-specific adapters enable targeted transformations, with pronounced benefits for structured enumeration styles (ingredients, dishware) over open-ended narratives.

Statistical Significance Analysis

To assess whether the performance differences between the proposed and baseline models reflect genuine improvements rather than sampling variance, paired bootstrap resampling tests ($N = 10,000$) were conducted on per-sample caption scores across both models. Figure 3 presents the mean difference ($\Delta = \text{Proposed} - \text{Baseline}$) with 95% confidence intervals across all five caption styles and five evaluation metrics.

The most statistically robust improvements occur in ingredient-focused captions, which yield significant gains across four metrics: CIDEr ($\Delta = 4.52$), BLEU-4 ($\Delta = 0.62$), ROUGE-L ($\Delta = 0.56$), and SPICE ($\Delta = 0.87$). This pattern confirms that the synergy between style-specific adapters and contrastive regularization is particularly effective for structured enumeration styles, where selective feature attention toward compositional and textural elements produces measurable and statistically reliable improvements.

For Detail+Aesthetic captions, ROUGE-L ($\Delta = 0.95$) and BLEU-4 ($\Delta = 0.76$) improvements reach statistical significance, representing the strongest lexical-level evidence in this study. These results validate that the contrastive module effectively separates aesthetic-oriented representations from other style clusters, enabling the model to generate more precise aesthetic language.

However, METEOR exhibits statistically significant degradation in two styles: Detail ($\Delta = -0.40$) and Detail+Aesthetic ($\Delta = -0.31$). Since METEOR captures synonym-level semantic alignment, this degradation is consistent with the trade-off observed in the quantitative results, where these open-ended styles involve higher linguistic variability that resists strict synonym matching. The remaining metrics across Short, Dishware, and Detail styles do not reach conventional significance thresholds ($p > 0.05$), reflecting wide per-sample variance inherent in these styles rather than an absence of improvement.

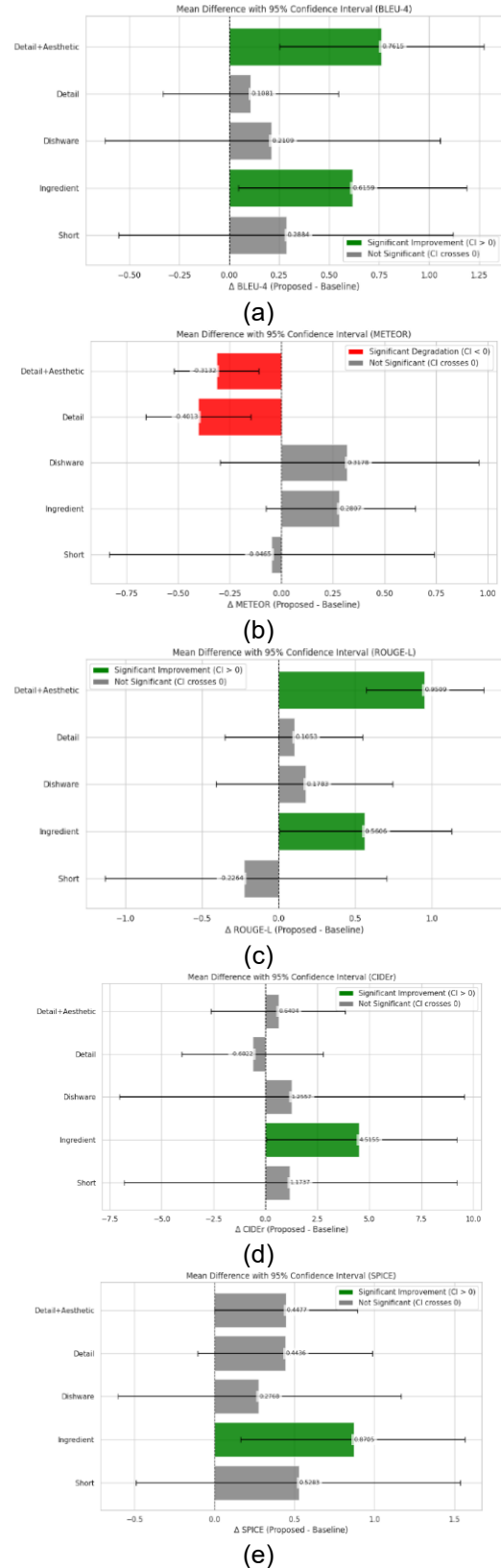


Figure 3. Mean difference ($\Delta = \text{Proposed} - \text{Baseline}$) with 95% confidence intervals from paired bootstrap resampling ($N = 10,000$) for (a) BLEU-4, (b) METEOR, (c) ROUGE-L, (d) CIDEr, and (e) SPICE across five caption styles.

Table 5. Ablation study on architectural component

Architecture	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Baseline	17.4	22.1	44.9	179.9	31.2
Baseline+Contrastive	17.3	21.6	44.5	175.9	30.9
Baseline+Style Adapter	17.1	21.7	44.6	177.3	31.0
Baseline+Both (Proposed)	17.5	21.9	45.2	181.4	31.8

Ablation Study

Table 5 reveals critical findings regarding architectural complementarity. Adding contrastive loss alone degrades performance (CIDEr: 175.9, -2.2%; SPICE: 30.9, -1.0%), confirming that contrastive objectives conflict with generative training without appropriate architectural support. This performance drop occurs because the Q-Former must simultaneously optimize two competing objectives: (1) contrastive loss pushes representations toward style-discriminative clustering, while (2) generative loss prioritizes semantic accuracy and fluency for caption generation. Without dedicated parametric capacity to handle these fundamentally different objectives, the model experiences conflicting optimization pressures that degrade overall performance. A metric-level inspection further supports this interpretation: METEOR exhibits the steepest relative decline (21.6 vs. 22.1, -2.3%), indicating that conflicting optimization pressures particularly disrupt semantic coherence in generation, as METEOR captures synonym-level semantic alignment. The concurrent drops in ROUGE-L (44.5 vs. 44.9, -0.9%) and BLEU-4 (17.3 vs. 17.4, -0.6%) confirm that this degradation is not isolated to a single evaluation dimension but reflects a systemic disruption of caption quality.

Similarly, style adapters alone yield minimal improvements (CIDEr: 177.3, -1.4%; SPICE: 31.0, -0.6%), likely due to representation drift where adapter outputs for different styles collapse into similar distributions. While adapters provide style-specific parametric capacity, they lack explicit mechanisms to enforce representational separation across styles, allowing their learned transformations to converge toward homogeneous representations that undermine style specialization. This is reflected in the metric pattern: while SPICE degrades only moderately (-0.6%), suggesting partial preservation of semantic relationships, BLEU-4 shows a steeper relative decline (17.1 vs. 17.4, -1.7%), indicating that unconstrained adapter transformations reduce lexical precision. Without geometric constraints from contrastive regularization, each adapter optimizes independently, allowing cross-style feature overlap that undermines n-gram consistency in caption generation.

Only the full model combining both components surpasses the baseline across multiple metrics: CIDEr (+0.8%), ROUGE-L (+0.7%), SPICE (+1.9%), and BLEU-4 (+0.6%), with only a marginal METEOR decrease (-0.9%), which is an acceptable trade-off in multi-style settings where diverse linguistic expression is preferable over strict synonym matching. The consistent improvement across structurally diverse metrics, including n-gram precision (BLEU-4), recall-oriented (ROUGE-L), and semantic graph (SPICE), confirms that the synergy between adapters and contrastive learning produces holistically better captions rather than optimizing for a single evaluation dimension.

This synergy has theoretical grounding in recent PEFT research on 3D pre-trained models [19], which demonstrated that adapter modules require complementary mechanisms to prevent representation drift. Their experiments showed that geometric-aware adapters alone were insufficient but combining them with regularization mechanisms yielded substantial improvements. The findings in this study extend this principle to multi-style captioning in instruction-tuned vision-language models. The contrastive module organizes the embedding space into style-aware clusters, stabilizing adapter learning through geometric constraints. A Study on multimodal fusion [23] supports this interpretation, noting that bottleneck mechanisms require proper regularization to avoid noise.

Comparison with Previous Study

No previous study directly addresses the specific combination explored in this study, namely multi-style food image captioning within an instruction-tuned vision-language framework using style-specific adapters and style-conditioned contrastive learning. The previous studies referenced here each contribute one dimension of this study independently under fundamentally different datasets, task formulations, and evaluation protocols, making direct numerical comparison inappropriate. Table 6 is therefore presented as a structured qualitative positioning rather than a performance benchmark.

Table 6. Comparison with previous study

Method	Task	Dataset	Instruction-aware	Adapter PEFT	Style Contrastive	CIDEr
InstructBLIP [14]	Image Captioning + VQA+Visual Reasoning, etc.	NoCaps, Flickr30K, etc.	Yes	No	No	123.1
VL-Adapter [18]	Image Captioning + VQA + Visual Reasoning, etc.	MSCOCO, VQAv2, GQA, etc.	No	Yes	No	114.9
SACO [24]	Image Captioning	PERSONALITY-CAP, SentiCap, FlickrStyle10K	No	No	Yes	21.0
OFA [25]	Image Captioning	Food-500 Cap	No	No	No	14.01
Baseline	Image Captioning	Food-500 Cap	Yes	No	No	179.9
Proposed	Image Captioning	Food-500 Cap	Yes	Yes	Yes	181.4

InstructBLIP achieves a NoCaps CIDEr of 123.1 and serves as the direct architectural foundation of this work. VL-Adapter establishes that adapter-based PEFT can match full fine-tuning using only 4.18% of parameters, achieving a MSCOCO CIDEr of 114.9. SACO demonstrates that style-aware contrastive objectives help models capture different visual content for each style, achieving a CIDEr of 21.0 on PERSONALITY-CAPTIONS. On Food-500 Cap, report that OFA achieves the highest zero-shot CIDEr of 14.01, confirming that general-purpose models without domain-specific fine-tuning fail to generate competent food captions. Despite sharing the same dataset, the absence of fine-tuning, style conditioning, and instruction awareness in OFA makes this result incomparable to those reported here.

The proposed method is the only approach that simultaneously incorporates instruction-awareness, adapter-based PEFT, and style-conditioned contrastive learning in the food domain, achieving CIDEr of 181.4 versus baseline 179.9 (+0.8%), ROUGE-L (+0.7%), and SPICE (+1.9%). These results confirm that integrating principles from InstructBLIP, VL-Adapter, and SACO into a unified framework yields meaningful improvements beyond what any single prior approach provides.

Collectively, the comparisons across sub-sections of this discussion demonstrate that the proposed method advances the state of the art along three complementary dimensions. In terms of parameter efficiency, it aligns with VL-Adapter [18] by achieving competitive gains without full fine-tuning. In terms of style discrimination, it extends the contrastive principle of SACO [24] into an instruction-tuned framework where style conditioning operates at the feature extraction stage rather than the output level. In terms of domain specificity, it substantially outperforms zero-shot general-purpose models such as OFA [25] (CIDEr 14.01) on Food-500 Cap, confirming that domain-specific instruction tuning with style-aware adaptation is essential for food image

captioning. Taken together, these results position the proposed method as a unified approach that simultaneously addresses instruction-awareness, style-specific adaptation, and contrastive regularization, a combination not previously explored in the context of multi-style food image captioning.

Qualitative Results

As reported in Table 7, qualitative analysis across two food image samples reveals nuanced improvements. For ingredient-focused captions in the first sample, the proposed model generates "toasted almonds" versus the baseline's "crushed almonds," demonstrating more nuanced characterization that reflects texture-based attributes. This aligns with observations from style-aware visual learning research [24], which found that different styles lead models to emphasize different visual aspects. In ingredient-focused descriptions, this means prioritizing textural and compositional properties over generic attributes. This pattern is further corroborated in the second sample (apple sauce), where the proposed model produces "light brown apple sauce" compared to the baseline's "dark brown apple sauce," demonstrating improved sensitivity to fine-grained color attributes. However, both models consistently miss hidden ingredients not prominently visible in the images, such as the crescent shape in the first sample and the cinnamon topping in the second, indicating remaining challenges in inferring compositional elements that are not visually salient.

The most notable improvement occurs in dishware descriptions, where the proposed model across both samples. In the first sample, the proposed model correctly identifies "blue and white patterned plate" while the baseline hallucinates "rustic wooden board," a critical factual error. Similarly, in the second sample, the proposed model accurately identifies "white bowl" while the baseline incorrectly describes "rustic ceramic bowl."

Table 7. Qualitative caption comparison across instruction styles

Image	Style	Architecture	Caption
	Short	Baseline	Crispy almond biscuits
		Proposed	Crispy almond biscuits
		Groundtruth	A pile of crispy almond biscuits
	Ingredients	Baseline	Crispy almond biscuits topped with crushed almonds
		Proposed	Crispy almond biscuits topped with toasted almonds
		Groundtruth	Crescent-shaped almond biscuits topped with visible almond slivers
	Disware	Baseline	Almond biscuits served on a rustic wooden board
		Proposed	Almond biscuits arranged on a blue and white patterned plate
		Groundtruth	Crispy almond biscuits served in a rustic wooden bowl
	Detail	Baseline	A plate of crispy almond biscuits sprinkled with almonds on the surface
		Proposed	A plate of crispy almond biscuits with some almonds on top
		Groundtruth	A pile of crescent-shaped crispy almond biscuits with lots of almonds clearly visible
	Detail +Aesthetic	Baseline	A plate of almond biscuits with a crispy texture and sprinkled with almonds. The soft lighting and gentle color tones enhance the inviting appearance of the biscuits, making them look fresh and appetizing.
		Proposed	Some almond biscuits on a blue and white plate. The soft lighting and subtle color tones enhance the inviting appearance of the biscuits, making them look fresh and appetizing.
		Groundtruth	A pile of crescent-shaped crispy almond biscuits with lots of almonds clearly visible. The soft lighting and gentle color tones create an inviting atmosphere, enhancing the natural appeal of the biscuits, making them look fresh and appetizing.
	Short	Baseline	A bowl of apple sauce
		Proposed	A bowl of apple sauce
		Groundtruth	A bowl of apple sauce
	Ingredients	Baseline	Dark brown apple sauce served with fresh green apples
		Proposed	Light brown apple sauce served with fresh green apples
		Groundtruth	Apple sauce with green apple pieces and cinnamon topping
	Disware	Baseline	Apple sauce served in a rustic ceramic bowl
		Proposed	Apple sauce served in a small white bowl on a pink napkin with three green apples beside it
		Groundtruth	Apple sauce served in a white bowl
	Detail	Baseline	A bowl of brown apple sauce with a spoon on a pink napkin with three green apples next to it
		Proposed	A bowl of brown apple sauce with three green apples next to it
		Groundtruth	A white bowl of apple sauce with visible green apple pieces and cinnamon topping
	Detail +Aesthetic	Baseline	A bowl of brown apple sauce with a spoon on a pink napkin with three green apples next to it. The soft lighting and gentle color tones enhance the inviting appearance of the dish, making it look fresh and appetizing.
		Proposed	A small bowl of apple sauce with green apples next to it. The soft lighting and gentle color tones enhance the natural appeal of the dish, making it look fresh and inviting
		Groundtruth	A white bowl of apple sauce with visible green apple pieces and cinnamon topping. The soft natural light and gentle color harmony enhance the refreshing appearance of the sauce, making it look inviting and fresh.

This consistent hallucination pattern in the baseline is directly traceable to the training data: as shown in Table 1, the GPT-4-generated training data contains hallucinated dishware descriptions, because GPT-4 received only text-based inputs without direct image access. The baseline model consequently learned and reproduced these hallucination patterns during inference. The proposed model's consistent

improvement over this pattern demonstrates that the bottleneck adapter architecture effectively mitigates training data bias by restricting information flow and filtering visually inconsistent features. This consistent pattern across both samples therefore strengthens the evidence that hallucination reduction is a systematic benefit of the proposed architecture rather than an isolated occurrence. Recent work on multimodal

understanding with bottleneck fusion [23] observed similar noise reduction properties, where limiting information exchange through bottleneck tokens helps prevent the propagation of irrelevant features that can lead to hallucinations.

For short captions, both models perform identically on simple single-entity descriptions, suggesting minimal benefit from style adaptation when the baseline already achieves near-optimal performance. Detailed captions show mixed results: the proposed model uses more concise phrasing in some cases while the baseline occasionally produces more descriptive output. In aesthetic-enhanced captions, both models successfully incorporate stylistic language such as "soft lighting" and "gentle color tones," though neither fully captures the specific aesthetic vocabulary of the ground truth, indicating that diversity in aesthetic expression remains a generalization challenge. Nevertheless, both models sometimes fail to detect fine-grained visual attributes such as the crescent shape and cinnamon topping, reflecting challenges when domain gaps between pre-training and target tasks are substantial [20].

Limitations

This study has several limitations that warrant acknowledgment. First, while the constructed dataset provides a strong foundation for multi-style training, GPT-4-generated captions introduce biases that affect model training. Since the dataset construction pipeline provides GPT-4 with only text-based inputs, namely the original captions and numerical aesthetic attributes, without direct access to the images, GPT-4 generates captions based on textual patterns rather than actual visual perception. Consequently, it occasionally produces factually incorrect descriptions. As illustrated in Table 1, the dishware caption for the pie image describes "a rustic wooden board" despite the image showing a white plate, demonstrating that GPT-4 hallucinates visual attributes when the original caption does not explicitly describe the serving vessel. Such training data bias can cause the model to learn and reproduce incorrect visual-style associations, ultimately affecting the quality of generated captions. Although the proposed bottleneck adapter architecture partially mitigates this bias, the generated captions do not fully resolve this issue because the model learns from incorrect style associations.

Second, the contrastive learning objective is constrained by the small batch size of 8, imposed by the memory requirements of the frozen EVA-CLIP ViT-g/14 and Vicuna-7B

components. With limited samples per batch, the number of available positive and negative pairs per style is insufficient to form well-separated style clusters in the embedding space, which may limit the full potential of the contrastive regularization mechanism.

CONCLUSION

This study introduces Style-Aware Contrastive Adapters for multi-style food image captioning, addressing the limitation of existing instruction-tuned models in generating stylistically differentiated captions. The proposed method integrates parallel style adapters and style-conditioned contrastive loss into InstructBLIP, achieving consistent improvements (CIDEr: 181.4 vs. baseline 179.9, +0.8%; SPICE: +1.9%), with substantial gains for structured caption styles (ingredient captions: +3.1% CIDEr).

The key finding is architectural complementarity: ablation studies confirm that contrastive loss alone degrades performance by 2.2% CIDEr, adapters alone by 1.4% CIDEr, while their synergistic combination surpasses the baseline. This validates that contrastive learning, when applied as a geometric regularizer, effectively stabilizes style-specific adaptations in instruction-aware image captioning. However, performance gains are inconsistent across instruction styles: structured styles such as ingredient and dishware benefit significantly (ingredient: +3.1% CIDEr, dishware: +0.4% CIDEr) because their descriptive patterns are well-defined and constrained, making them more amenable to style-specific adapter specialization and contrastive clustering. In contrast, detail-oriented captions show a trade-off between semantic correctness (SPICE +2.0%) and n-gram precision (CIDEr -0.6%), likely because these styles involve higher linguistic variability that is harder to regularize through the current adapter-contrastive learning combination. However, as detailed in the Limitations section, the current approach is constrained by global feature adaptation, training data biases from text-only GPT-4 generation, small batch size, and single-domain evaluation, all of which present opportunities for future improvement. Future research should explore visual grounding mechanisms, such as region-level attention or attribute-aware feature extraction, so that different styles attend to different visual regions rather than relying solely on global feature transformations; employing GPT-4 models with direct image access for dataset construction to ensure visually grounded and accurate style-specific captions; and adopting larger batch sizes through multi-GPU training to provide richer contrastive learning

signals, enabling the model to form more well-separated style representations in the embedding space.

ACKNOWLEDGMENT

The authors would like to thank the Lembaga Pengelola Dana Pendidikan (LPDP), the Ministry of Finance of the Republic of Indonesia, and the Department of Informatics, Faculty of Electrical Technology and Intelligent Informatics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia for supporting this research.

REFERENCES

- [1] M. Khairudin *et al.*, "Early detection of diabetes potential using cataract image processing approach," *Sinergi (Indonesia)*, vol. 28, no. 1, pp. 55–62, 2024, doi: 10.22441/sinergi.2024.1.006.
- [2] E. I. Anggraini *et al.*, "Development of face image recognition algorithm using CNN in airport security checkpoints for terrorist early detection," *Sinergi (Indonesia)*, vol. 29, no. 1, pp. 33–42, 2025, doi: 10.22441/sinergi.2025.1.004.
- [3] P. P. Khaing and M. T. Yu, "Two-Tier LSTM Model for Image Caption Generation," *Int. Journal of Intelligent Engineering and Systems*, vol. 14, no. 4, pp. 22–34, Aug. 2021, doi: 10.22266/ijies2021.0831.03.
- [4] F. Zhao, Z. Yu, T. Wang, and Y. Lv, "Image Captioning Based on Semantic Scenes," *Entropy*, vol. 26, no. 10, p. 876, Oct. 2024, doi: 10.3390/e26100876.
- [5] D. Abdal Hafeth and S. Kollias, "Insights into Object Semantics: Leveraging Transformer Networks for Advanced Image Captioning," *Sensors*, vol. 24, no. 6, p. 1796, Mar. 2024, doi: 10.3390/s24061796.
- [6] S. Rohitharun, L. Uday Kumar Reddy, and S. Sujana, "Image Captioning Using CNN and RNN," in *Proc. of the 2022 Asian Conference on Innovation in Technology (ASIANCON 2022)*, pp. 1–8, Aug. 2022. doi: 10.1109/ASIANCON55314.2022.9909146.
- [7] W. Xie *et al.*, "Recent advances in Transformer technology for agriculture: A comprehensive survey," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109412, 2024, doi: 10.1016/j.engappai.2024.109412.
- [8] S. Zhuang, P. Wang, G. Wang, D. Wang, J. Chen, and F. Gao, "Improving Remote Sensing Image Captioning by Combining Grid Features and Transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3135711.
- [9] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 18009–18019, doi: 10.1109/CVPR52688.2022.01750.
- [10] J. Zhang *et al.*, "BAE-ViT: An Efficient Multimodal Vision Transformer for Bone Age Estimation," *Tomography*, vol. 10, no. 12, pp. 2058–2072, Dec. 2024, doi: 10.3390/tomography10120146.
- [11] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *Proc. of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, vol. 162, July 2022, pp. 12888–12900.
- [12] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in *Proc. of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, vol. 202, Jul. 2023, pp. 19730–19742, doi: 10.5555/3618408.3619222.
- [13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in *Advances in Neural Information Processing Systems*, vol. 36, pp. 34892–34916, Dec. 2023.
- [14] W. Dai *et al.*, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," in *Advances in Neural Information Processing Systems*, vol. 36, pp. 49250–49267, Dec. 2023.
- [15] H. Xie *et al.*, "EmoViT: Revolutionizing Emotion Insights with Visual Instruction Tuning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26586–26595, Jun. 2024, doi: 10.1109/CVPR52733.2024.02511.
- [16] K. Basioti, M. A. Abdelsalam, F. Fancellu, V. Pavlovic, and A. Fazly, "CIC-BART-SSA: Controllable Image Captioning with Structured Semantic Augmentation," in *Computer Vision – ECCV 2024: 18th European Conference*, pp. 444–461, September 29–October 4, 2024. doi: 10.1007/978-3-031-72848-8_26.
- [17] Y. Niu, M. Zhu, and Z. Wei, "SamCap: Energy-based Controllable Image Captioning

- by Gradient-Based Sampling,” in *Proc. of the 2024 Int Conference on Multimedia Retrieval*, pp. 608–617, 2024. doi: 10.1145/3652583.3658112.
- [18] Y. L. Sung, J. Cho, and M. Bansal, “VL-ADAPTER: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks,” in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5217–5227, Jun. 2022, doi: 10.1109/CVPR52688.2022.00516.
- [19] Y. Tang et al., “Point-PEFT: Parameter-Efficient Fine-Tuning for 3D Pre-trained Models,” *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, pp. 5171–5179, Mar. 2024, doi: 10.1609/aaai.v38i6.28323.
- [20] I. Sel and D. Hanbay, “Efficient Adaptation: Enhancing Multilingual Models for Low-Resource Language Translation,” *Mathematics*, vol. 12, no. 19, p. 3149, Oct. 2024, doi: 10.3390/math12193149.
- [21] R. Wang et al., “K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1405–1418, Aug. 2021. doi: 10.18653/v1/2021.findings-acl.121.
- [22] R. Yuan, Y. Tang, Y. Wu, and W. Zhang, “Clustering Enhanced Multiplex Graph Contrastive Representation Learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 1341–1355, 2025, doi: 10.1109/TNNLS.2023.3334751.
- [23] P. Wang, X. Chen, Z. Shang, and W. Ke, “Multimodal Named Entity Recognition with Bottleneck Fusion and Contrastive Learning,” *IEICE Transactions on Information and Systems*, vol. E106D, no. 4, pp. 545–555, Apr. 2023, doi: 10.1587/TRANSINF.2022EDP7116.
- [24] Y. Zhou and G. Long, “Style-Aware Contrastive Learning for Multi-Style Image Captioning,” in *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2257–2267, May 2023. doi: 10.18653/v1/2023.findings-eacl.169.
- [25] Z. Ma et al., “Food-500 Cap: A Fine-Grained Food Caption Benchmark for Evaluating Vision-Language Models,” in *Proc. of the 31st ACM International Conference on Multimedia*, pp. 5674–5685, 2023. doi: 10.1145/3581783.3611994.
- [26] I. Kovács, M. Balázsne Lendvai, and J. Beke, “The Importance of Food Attributes and Motivational Factors for Purchasing Local Food Products: Segmentation of Young Local Food Consumers in Hungary,” *Sustainability*, vol. 14, no. 6, p. 3224, Mar. 2022, doi: 10.3390/su14063224.
- [27] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, “Photo Aesthetics Ranking Network with Attributes and Content Adaptation,” in *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, vol. 9905, pp. 662–679, 2016, doi: 10.1007/978-3-319-46448-0_40.