IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 June 2024, accepted 28 August 2024, date of publication 10 September 2024, date of current version 19 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3457586

RESEARCH ARTICLE

Architectural Synergies in Bi-Modal and Bi-Contrastive Learning

YUJIA GU[1], BRIAN LIU[2], TIANLONG ZHANG[3], XINYE SHA[4], AND SHIYONG CHEN[5]
[1]California State University at Long Beach, Long Beach, CA 90840, USA
[2]Stuyvesant High School, New York, NY 10282, USA
[3]University of Pittsburgh, Pittsburgh, PA 15260, USA
[4]Columbia University, New York, NY 10027, USA
[5]Beihang University, Beijing 100191, China

Corresponding author: Shiyong Chen (shiyongchen@buaa.edu.cn)

ABSTRACT The integration of visual and linguistic elements within artificial intelligence research is increasingly emphasized, spurred by advancements in pre-trained model technologies. Traditionally, such models have been developed independently, using methods like contrastive learning and image-captioning to boost their analytical and creative outputs. This paper introduces an innovative architecture known as the Zero-shot Unified Image-Text (ZsU-IT) framework, which synthesizes pre-training objectives into a cohesive Unicode-decoder structure. The ZsU-IT is intricately designed with distinct components for image and text processing, coupled with a bi-modal decoder, which seamlessly manages both encoding and decoding tasks across various functions. This dual functionality promotes an effective knowledge transfer between the visual and linguistic modalities, thereby enhancing the system's adaptability and efficiency in tasks like image-to-text translation and vice versa. Rigorous empirical studies reveal that ZsU-IT outstrips prevailing models across multiple applications, including image and text retrieval, image captioning, Visual Question Answering (VQA), and Stanford Natural Language Inference - Visual Entailment (SNLI-VE). This is particularly notable in complex settings involving sophisticated datasets such as medical texts and CT images. In zero-shot environments, ZsU-IT excels, displaying exceptional generalization capabilities. This prowess is highlighted by its significant achievements. The ZsU-IT framework not only sets a new benchmark in the fusion of vision and language technologies but also fosters novel opportunities for both ongoing research and practical implementations. This advancement marks a crucial step forward in the application of integrated multimodal data for complex problem-solving within the artificial intelligence landscape, paving the way for future breakthroughs.

INDEX TERMS Multimodal, domain adaption, visual-linguistic model.

I. INTRODUCTION

In the domain of integrating vision and language, modern approaches are typified by three predominant strategies. The first strategy widely adopted uses associated textual data alongside contrastive loss functions for the initial training of visual models. This method utilizes distinct unimodal encoders to process visual and textual inputs separately, thereby enhancing the models' abilities in classification tasks while facilitating zero-shot learning and the capacity for image-text retrieval [12].

Subsequent to this, the creation of generative models through pre-training configurations that merge image encoders with text decoders marks the second strategy. These systems apply Image-to-Text (I2T) token generation losses [20], proving effective in a variety of vision-language tasks, including Visual Question Answering and image captioning. Such configurations demonstrate their prowess in generating precise textual interpretations from visual inputs, underscoring their significance in handling complex multimodal challenges. The third strategic approach advances the methodology to include Text-to-Image (T2I) token generation losses. This technique utilizes models such as VQ-VAE and GANs to transform raw visual data into

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

discrete image tokens [4], [27], approaching the text-to-image generation as a sequence-to-sequence process. Renowned for its ability to produce intricately detailed images from textual descriptions, this method exemplifies the potent capabilities of sequential generative modeling in creating visual content from text.

In the rapidly evolving field of artificial intelligence, while many models focus on single objectives, a select few, such as CoCa [28], OFA [21], and UnifiedIO [11], ambitiously pursue dual goals. These models excel by utilizing a unified dataset of image-text pairs, endorsing the theory that a dual-objective framework can foster a more comprehensive understanding. The interplay between contrastive learning, which focuses on modality alignment, and generative tasks, which require detailed coordination between image and text, greatly benefits from such integrative strategies. Employing joint pre-training for these dual objectives not only deepens understanding but also enhances computational efficiency by facilitating shared frameworks throughout various phases of model training and deployment.

This study introduces a complex architectural framework called ZsU-IT. It effectively combines bi-modal and bi-contrastive learning methods, facilitating generative functions for both image-to-text and text-to-image transformations. Central to ZsU-IT are three essential components: an image unicoder, a text unicoder, and a cross-attention decoder. Drawing on the transformative potential of the Transformer architecture, these unicoders are crafted to fluidly transition between unimodal encoding and decoding roles. This flexibility is supported by uniquely tailored input embeddings and attention mechanisms specific to each mode of operation. In the contrastive learning phase, the unicoders serve as precise encoders, while during generative phases, they take on encoding tasks and participate in autoregressive decoding operations. The cross-attention decoder plays a crucial role by enhancing the synthesis of features from both visual and textual inputs, enabling a dynamic interchange of knowledge. This integration significantly enhances the effectiveness of the text-to-image (T2I) and image-to-text (I2T) conversion processes, merging the framework's varied pre-training objectives into a cohesive system.

This paper presents major advances and contributions to the fusion of vision and language modalities, outlined as follows:

- The introduction of an innovative integrated architecture, ZsU-IT, which harmoniously blends bimodal and bi-contrastive learning approaches with functionalities for both image-to-text and text-to-image transformations.
- Validation of the ZsU-IT model's efficacy through extensive training on a diverse dataset that includes both web-sourced image-text pairings and meticulously annotated high-quality image datasets.
- Demonstrations of significant improvements in the model's zero-shot learning capabilities and its performance in fine-tuned tasks, highlighting the robustness and versatility of the ZsU-IT framework.

II. LITERATURE SURVEY

A. ADVANCEMENTS IN VISUAL REPRESENTATION LEARNING THROUGH TEXTUAL SUPERVISION

The realm of visual representation learning has witnessed significant advancements due to the strategic utilization of textual data in pre-training visual processing frameworks. Prominent methodologies such as CLIP [14] and ALIGN [8] have capitalized on contrastive learning techniques to effectively map image-text pairs within an embedding space, distinctively segregating compatible and incompatible pairs. This approach has been instrumental in enhancing the models' capabilities in zero-shot visual recognition and improving the transferability of visual features. The boundaries of this field have been further expanded by developments in sophisticated models such as Florence [30], BASIC [13], and LiT [31], which significantly broaden the scope of the datasets and augment the computational prowess of these systems. In an innovative exploration, FILIP [25] delves into the refined application of local token features derived from both visual and textual sources, aiming to enhance the detail and effectiveness of contrastive learning. Concurrently, projects like MS-CLIP [26] and CLIPPO [19] are pioneering efforts to optimize the sharing of parameters across visual and textual modalities. These efforts are not only refining the efficiency of model training but are also setting new standards for how deeply integrated and efficient multimodal learning frameworks can be in handling complex, diverse datasets. This burgeoning research is continuously pushing the envelope, offering new perspectives and capabilities in the seamless integration of visual and textual data.

B. ADVANCED STRATEGIES IN PRE-TRAINING VISION-LANGUAGE MODELS

The merging of visual and linguistic elements through pre-training forms a vibrant field of study, showcasing tremendous promise for fostering intermodal comprehension. A particularly innovative approach in this area involves the use of a mask-reconstruction loss [22]. This method requires models to reconstruct inputs that are partially obscured, encompassing both image and text tokens, which promotes a deeper, more integrative understanding of the input data. Concurrently, auto-regressive techniques for generating text are utilized to enhance model training, yielding notable enhancements in performance [1]. These methods have been successfully implemented across various downstream applications, such as Visual Question Answering (VQA) [2] and automated image captioning. This progressive domain highlights the beneficial interaction between visual and textual data, leading to marked advancements in the model's ability to interpret and generate responses that closely mimic human interaction within complex multimodal contexts.

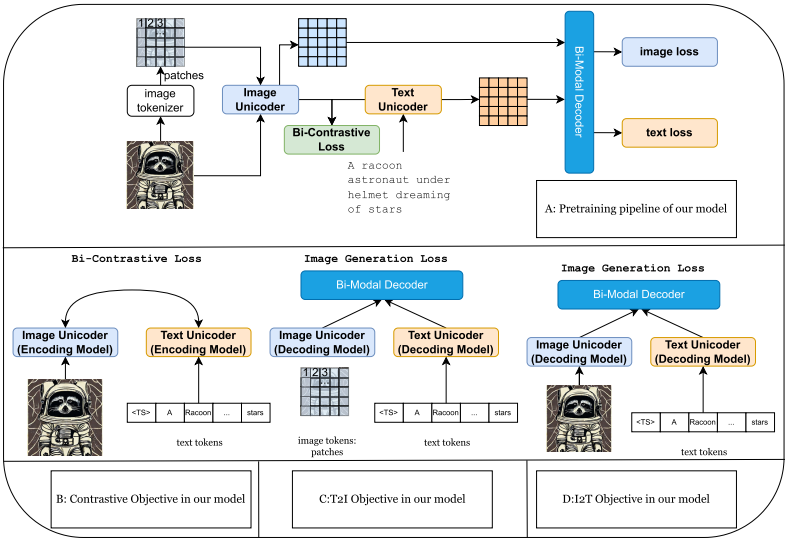

FIGURE 1. Summary of the ZsU-IT pre-training framework.

C. ADVANCES IN TEXT-TO-IMAGE SYNTHESIS

The field of text-to-image synthesis has emerged as a dynamic area of research within the broader machine learning community, marked by significant technological progress. The primary techniques employed in this domain can be broadly classified into diffusion and tokenization methods. Diffusion models [17] function by incrementally introducing noise into an image and then skillfully reversing this process based on textual inputs. This method allows for a controlled transformation guided by text descriptions, which effectively bridges the gap between linguistic content and visual representation. On the other hand, tokenization strategies involve the conversion of images into a series of discrete tokens using an image tokenizer, which are then reconstructed into coherent images using Transformer-based architectures. This reconstruction process leverages autoregressive mechanisms, akin to those used in machine translation [29], to sequentially predict image tokens. Alternatively, methods like those developed in MaskGIT [4] and Muse [3] employ simultaneous token prediction to enhance the efficiency and accuracy of image generation. These innovative approaches have propelled the capabilities of generative models, allowing for more precise and contextually relevant image synthesis from textual descriptions, thereby expanding the frontiers of creative artificial intelligence applications.

III. INVESTIGATING THE ZsU-IT FRAMEWORK: IN-DEPTH ANALYSIS

A. INPUT HANDLING PROCEDURES

The architectural ethos of the ZsU-IT model is grounded in its capacity for adaptability, engineered to accommodate a diverse spectrum of tasks. This flexibility is manifested in its ability to process three primary forms of input: textual data, discretized image tokens, and raw, unprocessed images. Each type of input undergoes a distinct preprocessing regimen, tailored to optimize the model's responsiveness and effectiveness, drawing upon well-established methodologies within the field [28].

1) TEXTUAL DATA PROCESSING

In handling textual data, the ZsU-IT model utilizes a SentencePiece tokenizer, which is configured with a comprehensive lexicon $\mathcal{L}$ consisting of 64,000 entries. This lexicon has been meticulously compiled from a broad sampling of datasets $\mathcal{D}$ specifically curated for pre-training:

$$\mathcal{L} = \bigcup_{d \in \mathcal{D}} \text{SentencePiece}(d, n = 64000) \tag{1}$$

where SentencePiece$(\cdot)$ denotes the SentencePiece tokenization function, and n represents the size of the lexicon. This strategy ensures that the model is equipped with a wide-ranging and representative linguistic base, priming it for robust performance across various textual scenarios.

2) DISCRETIZED IMAGE TOKENIZATION

For image processing, the ZsU-IT model employs an autoregressive mechanism essential for generating images from serialized image tokens [5], [6], [29]. This process leverages the established Parti methodology [29], utilizing a ViT-VQGAN tokenizer $\mathcal{T}$ that has been pre-trained and is maintained in a fixed state throughout the model's operations. Given an input image I, the tokenization process can be formulated as:

$$\mathbf{x} = \mathcal{T}(I) = [x_1, x_2, \ldots, x_T] \tag{2}$$

where $\mathbf{x}$ denotes the sequence of discrete image tokens, and T represents the length of the token sequence. This approach ensures that the transformation of two-dimensional images into a sequence of discrete tokens is both efficient and consistent with the model's generative objectives.

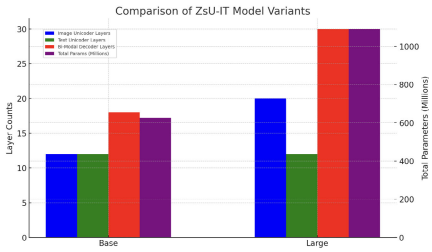

FIGURE 2. Dimensional varieties of ZsU-IT.

3) MANAGEMENT OF RAW IMAGE INPUTS

In addition to handling discrete image tokens, integrating raw images as inputs enhances the system's capabilities in tasks focused on image comprehension and the integration of image and text data. Let I_{raw} represent a raw input image. The model processes I_{raw} using a convolutional neural network (CNN) $f_{\text{CNN}}(\cdot)$ to extract visual features:

$$\mathbf{v} = f_{\text{CNN}}(I_{\text{raw}}) \tag{3}$$

where $\mathbf{v}$ denotes the extracted visual feature vector. This inclusion is essential for developing a more nuanced understanding of the visual and textual elements in tandem, thereby facilitating complex multimodal analysis.

B. CONSOLIDATED ARCHITECTURE SUMMARY

The structural composition of the Unified Image-Text (ZsU-IT) model is illustrated in Figure 1, which includes an image unicoder, a text unicoder, and a cross-modal attention decoder. The term "unicoder" emphasizes their dual capacity to function both as encoders and decoders, adapting their role depending on the task requirements. This multifaceted functionality is inspired by the work of Zhou et al. [32], which shows that a unified Transformer architecture can effectively manage bidirectional encoding for analysis-driven tasks and autoregressive decoding for creative outputs. When operating in a decoding role, the unicoders leverage previously encoded data to generate unimodal autoregressive outputs, forming a robust foundation for cross-modal generative applications. Ablation studies highlight the critical role of unicoders in enhancing the effectiveness of text-to-image (T2I) synthesis and in fostering a deeper multimodal understanding.

1) IMAGE UNICODER

The deployment of Vision Transformers (ViT) as a core technology in the image unicoder reflects their prominence in capturing detailed image features [9]. In this model, ViTs are uniquely adapted to serve not only as encoders but also as decoders for autoregressively generating image tokens. During the encoding phase, the image unicoder translates 2D image patches $\mathbf{p} = [p_1, p_2, \ldots, p_N]$ into a high-dimensional feature space using trainable linear projections:

$$\mathbf{z} = f_{\text{proj}}(\mathbf{p}) = [z_1, z_2, \ldots, z_N] \tag{4}$$

where $f_{\text{proj}}(\cdot)$ represents the linear projection function, and $\mathbf{z}$ denotes the sequence of projected patch features. These features undergo further refinement through several layers of Transformer architecture that apply bidirectional attention mechanisms:

$$\mathbf{h} = \text{Transformer}_{\text{enc}}(\mathbf{z}) \tag{5}$$

where $\mathbf{h}$ represents the output of the Transformer encoder layers. Conversely, during the decoding phase, the methodology shifts as images are tokenized into discrete tokens $\mathbf{x} = [x_1, x_2, \ldots, x_T]$ that are subsequently embedded:

$$\mathbf{e} = f_{\text{embed}}(\mathbf{x}) = [e_1, e_2, \ldots, e_T] \tag{6}$$

where $f_{\text{embed}}(\cdot)$ denotes the embedding function, and $\mathbf{e}$ represents the sequence of token embeddings. This Transformer architecture is then recalibrated for decoding purposes, incorporating causal, cone-shaped attention patterns that facilitate the sequential generation of image tokens [6], [29]:

$$\hat{\mathbf{x}} = \text{Transformer}_{\text{dec}}(\mathbf{e}) \tag{7}$$

where $\hat{\mathbf{x}}$ denotes the predicted image token sequence. This dual-mode functionality allows the unicoder to utilize shared parameters across different phases of operation, significantly enhancing the efficiency and quality of image generation from textual descriptions. This approach takes advantage of the integrated knowledge from the encoding phase, enriching the generative capabilities of the model far beyond conventional methods [28].

2) DYNAMICS OF THE TEXT UNICODER ARCHITECTURE

The text unicoder's architecture parallels its image-centric counterpart, endowed with dual roles in both encoding and decoding processes, achieved through the versatile use of Transformer architecture parameters. This strategic configuration allows for fluid switching between operational modes, utilizing a unified tokenizer and embedding layer that uniformly processes inputs and extracts token features efficiently:

$$\mathbf{e}_{\text{text}} = f_{\text{embed}}(\text{SentencePiece}(S)) \tag{8}$$

where S represents the input text sequence, and $\mathbf{e}_{\text{text}}$ denotes the resulting token embeddings. The decoding functionality is enhanced by a causal attention mask $\mathbf{M}_{\text{causal}}$, which sharpens the focus on sequentially significant data:

$$\hat{\mathbf{y}} = \text{Transformer}_{\text{dec}}(\mathbf{e}_{\text{text}}, \mathbf{M}_{\text{causal}}) \tag{9}$$

where $\hat{\mathbf{y}}$ represents the predicted text token sequence. Regarding text encoding techniques, the unicoder is adaptable to both bi-directional and causal attention mechanisms [25], [29], with studies showing minimal difference in their impact on performance. Consequently, causal attention has been standardized within our experimental setup for its straightforwardness and efficacy.

3) BI-MODAL DECODER: A MECHANISM FOR SYNERGISTIC GENERATION

Central to the generative prowess of our model, the bi-modal decoder integrates and transforms inputs from multiple modalities using a cross-attention framework [28]. It employs

auto-regressive text features $\mathbf{h}_{\text{text}}$ from the text unicoder in its decoding role to initiate text generation processes. Concurrently, it incorporates encoded visual data $\mathbf{h}_{\text{image}}$ as keys and values in the cross-attention layers, thereby enriching the textual outputs with pertinent visual details:

$$\hat{\mathbf{y}} = \text{CrossAttention}(\mathbf{h}_{\text{text}}, \mathbf{h}_{\text{image}}) \tag{10}$$

For image generation tasks, this process is reversed; auto-regressive image token features $\mathbf{h}_{\text{image}}$ from the image unicoder are refined using textual data $\mathbf{h}_{\text{text}}$ within the cross-attention setup:

$$\hat{\mathbf{x}} = \text{CrossAttention}(\mathbf{h}_{\text{image}}, \mathbf{h}_{\text{text}}, \mathbf{M}_{\text{cone}}) \tag{11}$$

where $\mathbf{M}_{\text{cone}}$ represents an innovative, cone-shaped masked sparse attention pattern used during image generation to minimize the computational demands typically associated with extensive sequence processing, a departure from the traditional causal attention used in text scenarios [16], [29].

In summary, the integrated design of unicoders and a bi-modal decoder within our model's architecture is meticulously tailored to excel in diverse multimodal understanding and generative tasks. The unicoders' flexible design, capable of toggling between encoding and decoding functionalities, ensures effective knowledge transfer across different stages of data processing. This comprehensive approach not only enhances the operational efficiency of the model but also significantly elevates its performance across a variety of multimodal tasks, demonstrating the robustness and adaptability of our system in pushing the boundaries of multimodal learning and generative technologies.

C. FRAMEWORKS FOR EARLY STAGE TRAINING

The pre-training strategy of the ZsU-IT model is carefully designed to fulfill three core objectives: optimizing a contrastive loss mechanism to correlate image and text data, and deploying specialized loss functions for both image-to-text (I2T) and text-to-image (T2I) conversions. The subsequent sections elaborate on these distinct loss functions and the methodologies implemented for effective scaling and initialization of the training process.

1) DUAL CONTRASTIVE LOSS

At the heart of the ZsU-IT pre-training regimen is the dual contrastive loss mechanism. This process involves the initial encoding of raw images and textual content through their dedicated unicoders, producing distinct encoded outputs $\mathbf{h}_{\text{image}}$ and $\mathbf{h}_{\text{text}}$ for each modality. For textual data, the encoding approach aligns with methodologies like those seen in CLIP [15] and ALIGN [8], where the feature vector of the terminal CLS token within the sequence serves as a holistic representation:

$$\mathbf{h}_{\text{text}} = \text{Transformer}_{\text{enc}}(\mathbf{e}_{\text{text}})[\text{CLS}] \tag{12}$$

Conversely, image features are extracted from a sequence of encoded vectors and are further refined through an innovative attention-based pooling mechanism, similar to the one proposed by Yu et al. [28]. This method employs a unique multi-head attention layer with a trainable query $\mathbf{q}$, which aggregates the features generated by the unicoder, treating them as both keys and values in the attention schema:

$$\mathbf{h}_{\text{image}} = \text{MultiHeadAttention}(\mathbf{q}, \mathbf{h}, \mathbf{h}) \tag{13}$$

This structured approach to feature integration is pivotal, as it allows for the subsequent application of a dual contrastive loss function. This function critically assesses and refines the alignment between corresponding image-text pairs while effectively distinguishing them from mismatched pairs in the same batch. By doing so, it significantly bolsters the model's ability to discern and link relevant features across visual and textual modalities, thereby enhancing the overall efficacy of the multimodal learning process. The dual contrastive loss $\mathcal{L}_{\text{Con}}$ is formulated as follows:

$$\mathcal{L}_{\text{Con(text2image)}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\frac{\mathbf{h}_{\text{text}_i}^{\top}\mathbf{h}_{\text{image}_i}}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\mathbf{h}_{\text{text}_i}^{\top}\mathbf{h}_{\text{image}_j}}{\tau}\right)},$$

$$\mathcal{L}_{\text{Con(image2text)}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\frac{\mathbf{h}_{\text{image}_i}^{\top}\mathbf{h}_{\text{text}_i}}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\mathbf{h}_{\text{image}_i}^{\top}\mathbf{h}_{\text{text}_j}}{\tau}\right)},$$

$$\mathcal{L}_{\text{Con}} = \mathcal{L}_{\text{Con(text2image)}} + \mathcal{L}_{\text{Con(image2text)}}, \tag{14}$$

where N denotes the batch size, and τ represents a temperature parameter that controls the sharpness of the similarity distribution. This dual contrastive loss not only promotes a deeper understanding between the linked modalities but also serves as a foundational element in the model's training architecture, driving the development of a robust framework capable of advanced multimodal integration.

2) QUANTITATIVE EVALUATION OF GENERATION LOSS IN I2T AND T2I TASKS

The optimization of I2T and T2I tasks employs a cross-entropy loss function integrated into the bi-modal decoder. This loss function is crucial for enhancing the probability of accurately predicting subsequent tokens in a sequence, which aligns with the auto-regressive modeling strategy. This approach facilitates sequential data processing by predicting each subsequent token based on the previously generated context, thereby improving the overall coherence and accuracy of the generated outputs. The specific mathematical formulations for the I2T and T2I generation losses are presented below, highlighting the detailed mechanics of how these losses are computed and applied within our model's framework.

For the image-to-text (I2T) task, given an input image I and the ground-truth text sequence $\mathbf{y} = [y_1, y_2, \ldots, y_T]$, the I2T generation loss $\mathcal{L}_{\text{I2T}}$ is computed as:

$$\mathcal{L}_{\text{I2T}} = -\sum_{t=1}^{T}\log P_\theta\left(y_t | y_1, \ldots, y_{t-1}, I\right) \tag{15}$$

where P_θ denotes the conditional probability distribution modeled by the bi-modal decoder, parameterized by θ. This loss function quantifies the discrepancy between the predicted text tokens and the ground-truth tokens, conditioned onthe input image and the preceding textual context.

Similarly, for the text-to-image (T2I) task, given an input text sequence $\mathbf{y}$ and the ground-truth image token sequence $\mathbf{x} = [x_1, x_2, \ldots, x_T]$, the T2I generation loss $\mathcal{L}_{\text{T2I}}$ is computed as:

$$\mathcal{L}_{\text{T2I}} = -\sum_{t=1}^{T} \log P_\theta \left(x_t | x_1, \ldots, x_{t-1}, \mathbf{y}\right) \tag{16}$$

In this formulation, the bi-modal decoder predicts each image token based on the input text sequence and the previously generated image tokens.

The variables defined in these equations carry specific meanings essential for understanding the computation of generation losses in our model:

- y_t and x_t represent the textual and image tokens at position t within their respective sequences, critical for sequential generation tasks.
- $y_1, \ldots, y_{t-1}$ and $x_1, \ldots, x_{t-1}$ denote the sequences of textual and image tokens that provide the necessary context leading up to the t^{th} token, thereby facilitating contextually aware generation.
- I and $\mathbf{y}$ are the inputs for the image and text, respectively, used to condition the model in the I2T and T2I tasks, ensuring that the generation is relevant and accurately aligned with the provided inputs.

These elements play crucial roles in the mathematical framework of our model, ensuring accurate and context-aware generation in both I2T and T2I tasks. This methodical strategy for delineating and enhancing the Image-to-Text (I2T) and Text-to-Image (T2I) generation losses is a critical element of the ZsU-IT model's pre-training protocol. The approach is carefully crafted to not only improve the model's ability to perform these specific generative tasks effectively but also to enrich its ability to interpret and navigate the complex interactions between visual and textual modalities.

To further analyze the properties and effectiveness of the I2T and T2I generation losses, we can consider their gradients with respect to the model parameters θ.

Furthermore, we can analyze the convergence properties of the generation losses by examining their behavior during training. Let θ_k denote the model parameters at training iteration k. The update rule for the parameters using a gradient descent optimizer with learning rate α is given by:

$$\theta_{k+1} = \theta_k - \alpha \nabla_\theta \mathcal{L}(\theta_k) \tag{17}$$

where $\mathcal{L}$ represents either $\mathcal{L}_{\text{I2T}}$ or $\mathcal{L}_{\text{T2I}}$, depending on the task. As the training progresses and the model parameters are updated iteratively, the generation losses are expected to decrease, indicating an improvement in the model's ability to generate accurate and coherent outputs.

The convergence of the generation losses can be formally analyzed using techniques from optimization theory, such as the convergence analysis of stochastic gradient descent (SGD). Under certain assumptions, such as the Lipschitz continuity of the loss function and the bounded variance of the stochastic gradients, it can be shown that the expected value of the generation losses decreases over the course of training, guaranteeing convergence to a local minimum.

Mathematically, let $\mathcal{L}^*$ denote the optimal value of the loss function, and let $\mathbb{E}[\cdot]$ represent the expectation operator. The convergence property can be expressed as:

$$\mathbb{E}[\mathcal{L}(\theta_k)] - \mathcal{L}^* \leq \frac{C}{\sqrt{k}} \tag{18}$$

where C is a constant that depends on the Lipschitz constant of the loss function and the variance of the stochastic gradients. This inequality indicates that the expected value of the generation loss approaches the optimal value at a rate of $\mathcal{O}(1/\sqrt{k})$ as the number of training iterations k increases.

In practice, the convergence of the generation losses is monitored using validation metrics evaluated on a held-out dataset. Early stopping techniques can be employed to prevent overfitting and ensure that the model generalizes well to unseen data.

By optimizing these generation losses, the ZsU-IT framework helps to foster a more nuanced understanding and processing capability, thereby enhancing the overall functionality and adaptability of the model in handling diverse multimodal scenarios. The mathematical analysis of the loss functions, their gradients, and convergence properties provides a theoretical foundation for the effectiveness of the pre-training protocol and its ability to generate high-quality outputs in both I2T and T2I tasks.

D. INTEGRATED ADVERSARIAL DOMAIN ADJUSTMENT

In addressing the distribution discrepancies between source and target domains, we integrate the Cross-Domain Adaptive Approach (CADA) [23] into our model's training framework. This innovative method is specifically engineered to bridge the disparities between the well-labeled source domain $\mathcal{D}^s$ and the unlabeled target domain $\mathcal{D}^t$, enhancing the model's performance through strategic feature and structure transfer. Unlike earlier strategies that often compromised discriminative structures for domain alignment, our approach aims to simultaneously reduce disparities across the combined distributions.

Let $P^s(\mathbf{x}^s, y^s)$ and $P^t(\mathbf{x}^t, y^t)$ denote the joint distributions of the source and target domains, respectively, where $\mathbf{x}^s$ and $\mathbf{x}^t$ represent the input features, and y^s and y^t represent the corresponding labels. Given the complexities associated with modeling the joint distributions directly, our method introduces a refined technique for estimating the differences in these distributions. This is achieved through a calculated analysis of the features and structural elements present in each domain. By doing so, we provide a more nuanced and effective framework for domain adaptation, which addresses disparities on both the feature level and the structural level. This integrated adversarial domain adjustment strategy not

only enhances the adaptability of our model but also ensures a more robust generalization across diverse domain settings.

1) PERIPHERAL ADVERSARIAL DOMAIN ADJUSTMENT

Within the realm of domain adversarial training, the Peripheral Adversarial Domain Adjustment, grounded in the foundational work of Ganin and Lempitsky on marginal adversarial loss $\mathcal{L}_{\text{madv}}$, plays a crucial role in our approach. This concept is pivotal in mitigating domain discrepancies, particularly in applications like GestureQuery, where it effectively facilitates domain adaptation by promoting robust feature discrimination that is invariant to domain-specific distortions.

Let G denote the feature extractor network, and let D represent the domain discriminator network. The objective of the marginal adversarial loss is to minimize the divergence between the feature distributions of the source and target domains, which can be formally expressed as:

$$\mathcal{L}_{\text{madv}} = \mathbb{E}_{\mathbf{x}^s \sim \mathcal{D}^s}\left[\log D(G(\mathbf{x}^s))\right] + \mathbb{E}_{\mathbf{x}^t \sim \mathcal{D}^t}\left[\log(1 - D(G(\mathbf{x}^t)))\right] \tag{19}$$

By minimizing this loss, the feature extractor G learns to generate domain-invariant features, while the domain discriminator D attempts to distinguish between the source and target domains based on these features. The adversarial training process encourages the feature extractor to produce representations that fool the domain discriminator, thereby aligning the marginal feature distributions across domains.

2) RELATIONAL ADVERSARIAL DOMAIN ADJUSTMENT

Expanding upon traditional domain adaptation strategies that focus on matrix optimization [10], our framework advances a novel method for evaluating the divergence in conditional distributions between domains. This approach is encapsulated in the following equation:

$$\left|P^s\left(y^s \mid \mathbf{x}^s\right) - P^t\left(y^t \mid \mathbf{x}^t\right)\right| \propto \left|P^s\left(\mathbf{x}^s \mid y^s\right) - P^t\left(\mathbf{x}^t \mid y^t\right)\right| \tag{20}$$

This equation highlights our method's focus on minimizing the disparities between conditional distributions across domains. The relational adversarial domain adjustment strategy addresses this by establishing a mathematical relation between the conditional distributions of source and target domains, which helps in refining the domain adaptation process.

Let D_k denote the class-specific domain discriminator for the k^{th} class. The relational adversarial loss $\mathcal{L}_{\text{dadv}}$ is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{dadv}} = &-\sum_{k=1}^{C} \mathbb{E}_{\mathbf{x}_i^{s,k} \sim \mathcal{D}^{s,k}} \log D_k\left(G\left(\mathbf{x}_i^{s,k}\right)\right) \\ &- \mathbb{E}_{\mathbf{x}_i^{t,k} \sim \mathcal{D}^t} \log\left(1 - D_k\left(G\left(\mathbf{x}_i^{t,k}\right)\right)\right)\end{aligned} \tag{21}$$

where $\mathbf{x}^{s/t,k}$ denotes data instances that belong to the k^{th} class from either the source or target domain, and C represents the total number of classes. By minimizing this loss, the model learns to align the conditional feature distributions between the source and target domains for each class, thereby enhancing its ability to generalize across domains.

Calculating the divergence in Eq. 20 using both authentic and synthetic labels in the context of deep neural networks presents significant challenges, primarily due to the intricacies involved in batch sampling and model training dynamics. Our approach aims to tackle these complexities to enhance model generalization across varied domain environments.

3) GEOMETRIC GRAPH STRUCTURING

To advance our sampling methodology, we describe a detailed procedure for creating instance relationship graphs for both the source and target domains. This process begins with ensuring parity between mini-batch samples, denoted as $B^{s/t}$, for both domains. At the activation map layer l, represented as $l\mathbb{G}^{s/t,l} \in \mathbb{R}^{B^{s/t} \times K^{s/t} \times H^{s/t} \times W^{s/t}}$, the feature map $\mathbb{G}^{s/t,l}$ is transformed into a reshaped matrix $\mathbf{A}^{s/t,l} \in \mathbb{R}^{B^{s/t} \times (C^{s/t} H^{s/t} W^{s/t})}$.

The Gram Matrix $\mathbf{Q}^{s/t,l}$, which captures the internal instance relationships within each domain, is computed as:

$$\mathbf{Q}^{s/t,l} = \mathbf{A}^{s/t,l}(\mathbf{A}^{s/t,l})^{\top} \tag{22}$$

where $\mathbf{A}_{i\bullet}^{s/t,l}$ and $\mathbf{A}_{j\bullet}^{s/t,l}$ denote the activation maps for instances i and j, respectively. These inner product calculations are critical for building the final instance relationship graph, denoted as $\mathbf{R}^{s/t,l}$. To enhance the clarity and effectiveness of this representation, each row within $\mathbf{Q}_{i\bullet}^{s/t,l}$ undergoes L_2 normalization. This normalization standardizes the relationship representations, ensuring a precise and effective mapping of instance relationships, which is vital for understanding and modeling the interactions within and between domains.

The geometric graph matching loss $\mathcal{L}_{\text{tgm}}$, which quantitatively assesses the disparity between instance relationship graphs derived from the source and target domains, is defined as:

$$\mathcal{L}_{\text{tgm}} = \sum \frac{1}{B^{s/t}} \left\| \mathbf{R}^{s,l}\left(G\left(\mathbf{x}^s\right)\right) - \mathbf{R}^{t,l}\left(G\left(\mathbf{x}^t\right)\right) \right\|_2^2 \tag{23}$$

where $\mathbf{R}^{s,l}(G(\mathbf{x}^s))$ and $\mathbf{R}^{t,l}(G(\mathbf{x}^t))$ denote the instance relationship graphs for the source and target domains, respectively, derived from the activations produced by the neural network function G.

Optimizing $\mathcal{L}_{\text{tgm}}$ during the training process is crucial for aligning the structural characteristics of data between the two domains. By minimizing this loss, the model effectively reduces the feature space discrepancies that typically hinder effective domain adaptation, particularly in cross-domain classification tasks. This alignment ensures that the model not only recognizes and processes features universally across domains but also retains the ability to discern domain-specific nuances essential for accurate classification.

To further analyze the convergence properties of $\mathcal{L}_{\text{tgm}}$, we consider the update rule for the model parameters θ using

a gradient descent optimizer with learning rate α:

$$\theta_{k+1} = \theta_k - \alpha \nabla_\theta \mathcal{L}_{\text{tgm}}(\theta_k) \tag{24}$$

where k denotes the iteration index. Under suitable assumptions, such as the Lipschitz continuity of the loss function and the bounded variance of the stochastic gradients, it can be shown that the expected value of $\mathcal{L}_{\text{tgm}}$ decreases over the course of training, guaranteeing convergence to a local minimum [3]:

$$\mathbb{E}[\mathcal{L}_{\text{tgm}}(\theta_k)] - \mathcal{L}^*_{\text{tgm}} \leq \frac{C}{\sqrt{k}} \tag{25}$$

where $\mathcal{L}^*_{\text{tgm}}$ denotes the optimal value of the loss function, and C is a constant that depends on the Lipschitz constant of the loss function and the variance of the stochastic gradients. This inequality indicates that the expected value of $\mathcal{L}_{\text{tgm}}$ approaches the optimal value at a rate of $\mathcal{O}(1/\sqrt{k})$ as the number of training iterations k increases.

Thus, the minimization of $\mathcal{L}_{\text{tgm}}$ facilitates a more unified and coherent representation of features across domains, significantly enhancing the model's ability to perform reliably in cross-domain scenarios. This strategic alignment is instrumental in boosting the overall efficacy and adaptability of the model in domain transfer and generalization tasks, crucial for successful real-world applications where domain variance is prevalent.

E. ENHANCEMENTS THROUGH CLASSIFIER-FREE STEERING IN TEXT-TO-IMAGE SYNTHESIS

Incorporating classifier-free guidance (CFG) [7] into our text-to-image (T2I) synthesis framework enhances the accuracy and relevance of the generated images concerning their textual descriptions. During the model training phase, input text tokens, which form the conditioning vectors, are randomly masked with a probability of 10%. This mechanism facilitates a dual-prediction approach during the inference stage: one prediction uses the original, unmasked text tokens, represented as $I(z, T)$, and the other uses entirely masked text inputs, denoted as $I(z)$. These predictions are then integrated using linear interpolation to compute the final image output as follows:

$$I = I(z) + \alpha(I(z, T) - I(z)) \tag{26}$$

where α is a modifiable parameter controlling the extent of influence by the classifier-free guidance, set at 2.0 for our experiments.

To analyze the effect of CFG on the generated images, we consider the gradient of the output image I with respect to the latent variable z:

$$\nabla_z I = \nabla_z I(z) + \alpha(\nabla_z I(z, T) - \nabla_z I(z)) \tag{27}$$

The gradient $\nabla_z I(z, T)$ represents the direction in the latent space that maximizes the alignment between the generated image and the conditioned text T, while $\nabla_z I(z)$ represents the unconditioned gradient. By incorporating CFG, the model effectively balances the influence of the conditioned and unconditioned gradients, allowing for a more controlled and text-aligned image generation process.

1) HOLISTIC LOSS FORMULATION

Our pre-training framework integrates three distinct loss functions, aiming to optimize the model's capabilities in both generative (text-to-image and image-to-text) and interpretive aspects. This integrated loss function promotes the synergistic improvement of the model's ability to generate images from text and vice versa. The role of classifier-free guidance in this setup is crucial as it refines the model's performance in T2I tasks by adeptly manipulating the conditioning vectors. The composite loss function is expressed as:

$$\mathcal{L}_{\text{ZsU-IT}} = \lambda_{\text{Con}}\mathcal{L}_{\text{Con}} + \lambda_{\text{I2T}}\mathcal{L}_{\text{I2T}} + \lambda_{\text{T2I}}\mathcal{L}_{\text{T2I}} \tag{28}$$

This structured approach strategically enhances the model's interpretive and generative efficiencies by integrating detailed textual and visual understandings within a singular training regimen.

To analyze the convergence properties of the holistic loss $\mathcal{L}_{\text{ZsU-IT}}$, we consider the update rule for the model parameters θ using a gradient descent optimizer with learning rate α:

$$\theta_{k+1} = \theta_k - \alpha \nabla_\theta \mathcal{L}_{\text{ZsU-IT}}(\theta_k) \tag{29}$$

Under suitable assumptions, such as the Lipschitz continuity of the loss function and the bounded variance of the stochastic gradients, it can be shown that the expected value of $\mathcal{L}_{\text{ZsU-IT}}$ decreases over the course of training, guaranteeing convergence to a local minimum [3]:

$$\mathbb{E}[\mathcal{L}_{\text{ZsU-IT}}(\theta_k)] - \mathcal{L}^*_{\text{ZsU-IT}} \leq \frac{C}{\sqrt{k}} \tag{30}$$

where $\mathcal{L}^*_{\text{ZsU-IT}}$ denotes the optimal value of the loss function, and C is a constant that depends on the Lipschitz constants of the individual loss components and the variance of the stochastic gradients. This inequality indicates that the expected value of $\mathcal{L}_{\text{ZsU-IT}}$ approaches the optimal value at a rate of $\mathcal{O}(1/\sqrt{k})$ as the number of training iterations k increases.

To further analyze the dynamics of the holistic loss optimization, we consider the gradients of the individual loss components with respect to the model parameters θ:

$$\nabla_\theta \mathcal{L}_{\text{Con}} = \nabla_\theta \mathcal{L}_{\text{Con(text2image)}} + \nabla_\theta \mathcal{L}_{\text{Con(image2text)}} \tag{31}$$

These gradients guide the optimization process, adjusting the model parameters to minimize the respective loss components. The contrastive loss gradient $\nabla_\theta \mathcal{L}_{\text{Con}}$ encourages the model to learn discriminative features that align corresponding image-text pairs while distinguishing them from mismatched pairs. The I2T and T2I loss gradients, $\nabla_\theta \mathcal{L}_{\text{I2T}}$ and $\nabla_\theta \mathcal{L}_{\text{T2I}}$, guide the model to generate accurate and contextually relevant outputs conditioned on the input modalities.

The holistic loss formulation, along with the adaptive weighting scheme and the classifier-free guidance, enables the ZsU-IT model to effectively capture the complex interactions between visual and textual modalities. By optimizing

this composite loss function, the model learns to generate high-quality images from textual descriptions and accurate textual descriptions from images, while also developing a deep understanding of the semantic relationships between the two modalities.

The convergence analysis and the study of the gradient dynamics provide theoretical insights into the effectiveness of the ZsU-IT pre-training framework. The mathematical formulations and the optimization techniques employed in this framework contribute to its robustness, adaptability, and generalization capabilities, making it a powerful tool for multimodal learning tasks.

In conclusion, the integrated adversarial domain adjustment, classifier-free guidance, and holistic loss formulation constitute the core components of the ZsU-IT pre-training framework. These techniques work synergistically to enhance the model's ability to bridge the gap between visual and textual modalities, enabling it to generate high-quality outputs and develop a deep understanding of the semantic relationships between the two modalities. The mathematical analysis of these components provides a solid theoretical foundation for the effectiveness of the ZsU-IT model and its potential for real-world applications in various domains, such as computer vision, natural language processing, and multimedia analysis.

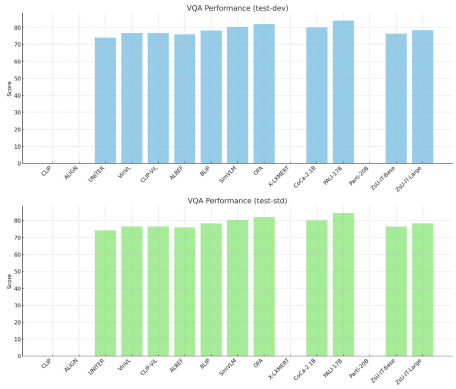

FIGURE 3. Analysis of diverse models using randomly altered data.

IV. ANALYTICAL PERSPECTIVES ON THE INITIAL TRAINING OF ZSU-IT

This section presents an exhaustive overview of the ZsU-IT model's pre-training regimen, detailing the foundational datasets, optimization techniques, and subsequent fine-tuning protocols to offer an extensive understanding of the pre-training environment (refer to Section IV-A). Following segments, Sections IV-B and IV-C, delve into the empirical assessments conducted in zero-shot frameworks and through fine-tuning across a range of tasks, respectively. These evaluations cover three critical areas: (1) visual comprehension, (2) image captioning integrated with multimodal understanding, and (3) text-to-image content generation. Detailed ablation studies further probe the architectural rationale behind the ZsU-IT model, substantiating the design choices implemented.

A. GUIDE TO PRE-TRAINING PRACTICES

In the context of T2I generation, the model processes 2,048 image-text pairings from the ALIGN and WebLI datasets. The ZsU-IT models are subjected to a stringent pre-training regimen, encompassing 2 million steps, and utilizes the Adafactor optimizer with a weight decay factor of 0.055. A gradual warm-up phase initially raises the learning rate to $4.5e-5$ over the first 3,000 steps, followed by an exponential decay starting at step 80,300. This comprehensive pre-training phase spans approximately 10 days, utilizing the 8 GTX A100 80GB GPUs.

Additionally, a secondary fine-tuning phase is implemented to enhance the model's ability to discern finer visual nuances. This phase utilizes high-resolution raw images, specifically adjusted to 578×578 pixels, to deepen the image encoding processes. Spanning 40,000 steps, this fine-tuning phase is crucial for adapting ZsU-IT to capture more subtle and sophisticated visual features, thus improving its performance across various downstream applications.

Together, the pre-training and fine-tuning strategies of the ZsU-IT model are intricately designed to leverage the combined strengths of diverse datasets and optimized loss functions. This comprehensive approach not only fosters robust visual and linguistic capabilities within ZsU-IT but also prepares it to excel in specific task-driven applications, highlighting a dedication to advancing the field of multimodal machine learning.

B. REVIEW OF ZERO-SHOT PERFORMANCE IN APPLICATIONS

The ZsU-IT framework represents a paradigm shift in the utilization of deep learning for multimodal tasks, demonstrating marked proficiency in a range of zero-shot applications. Notably, the ZsU-IT-Large variant has established new performance benchmarks in zero-shot image classification, recording a remarkable accuracy of 82.7% on the ImageNet dataset. This achievement exceeds the outputs of renowned predecessors such as CLIP and ALIGN, highlighting the efficacy of ZsU-IT's comprehensive pre-training regimen. This regimen skillfully integrates textual and visual data, catalyzing the formation of powerful and adaptable representational models suitable for diverse modalities.

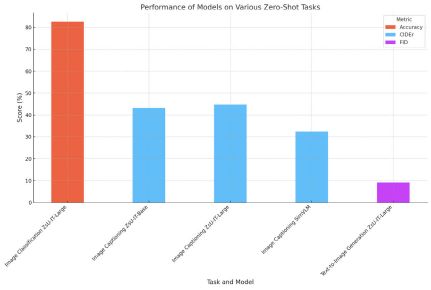

FIGURE 4. Outcome analysis across multiple zero-shot scenarios.

In zero-shot image-text retrieval, ZsU-IT-Large has surpassed existing benchmarks on prominent datasets like Flickr and MS-COCO, where it excelled in 5 of the 8 metrics

assessed. This performance underscores its ability to accurately correlate images with their corresponding textual descriptions, a fundamental skill for sophisticated retrieval systems. Additionally, ZsU-IT has made significant strides in zero-shot image captioning, particularly demonstrated by its performance on the MS-COCO dataset. Here, the ZsU-IT-Base and ZsU-IT-Large variants achieved CIDEr scores of 43.0 and 44.8, respectively, significantly surpassing the SimVLM model by substantial margins. This performance emphasizes ZsU-IT's capacity to generate contextually precise and linguistically coherent image captions.

Furthermore, ZsU-IT-Large extends its capabilities to zero-shot text-to-image generation, achieving a FID score of 9.37 on the MS-COCO dataset. This score not only outperforms larger models such as DALL-E 2 and Make-A-Scene but also confirms ZsU-IT's superior ability in producing visually appealing and contextually appropriate images from textual prompts. This illustrates a profound grasp of the complex interplay between textual narratives and visual outputs.

The collective accomplishments of ZsU-IT underscore its critical impact on advancing AI research, particularly in the domain of multimodal learning. The model's stellar performance across various zero-shot tasks opens up promising prospects for diverse applications, including enhanced image and video analysis, advanced natural language processing techniques, and pioneering methodologies in robotics. The integration of multiple pre-training strategies within a unified framework by ZsU-IT highlights the transformative potential of AI to comprehend and generate human-like responses across different modalities, paving the way for significant advancements in the field.

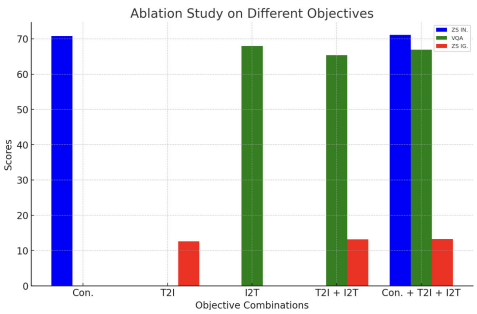

FIGURE 5. Objective reduction study: 'con.' represents contrastive loss.

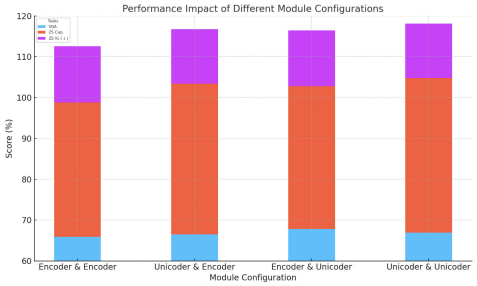

FIGURE 6. Comparative ablation: unicoder versus encoder.

C. ENHANCEMENTS THROUGH APPLICATION-SPECIFIC TUNING

To demonstrate the versatility of the Universal Image Transformer (ZsU-IT) across a broad range of applications including image understanding, multimodal integration, and text-driven visual synthesis, we employed both linear probing and in-depth fine-tuning methods across various secondary tasks. In the specific case of linear probing on the ImageNet dataset, we fixed all parameters within the image unicoder, focusing solely on optimizing a linear classifier for image categorization tasks. This approach underlined the ZsU-IT-Large model's superiority, where it outperformed established models like CLIP and ALIGN by approximately 1%.

In domains requiring synergy between visual and textual data, we explored tasks such as VQA, Stanford Natural Language Inference-Visual Entailment (SNLI-VE), and image captioning. Each of these tasks demands a thorough comprehension of both visual elements and textual narratives. The ZsU-IT model was subjected to extensive fine-tuning across all its parameters, with its effectiveness gauged against validation and test datasets. Specifically in the context of image captioning, ZsU-IT employed the same computational strategy as in its zero-shot captioning capabilities, as outlined in Figure 3, achieving impressive scores in the Consensus-based Image Description Evaluation (CIDEr) metric without resorting to task-specific optimizations such as CIDEr tuning. For fairness in evaluation, results were uniformly derived using the straightforward cross-entropy loss metric.

The outcomes from these application-specific tuning exercises verify the robustness of ZsU-IT's pre-trained representations across a diverse suite of tasks, positioning it as a leader among current models in various performance assessments. This not only highlights the model's expansive application potential and effectiveness within our foundational training framework but also underscores its prospective utility across multiple domains including visual and video analytics, computational linguistics, and robotics. This broad applicability suggests promising directions for future expansions and applications of the ZsU-IT framework in advancing multimodal machine learning technologies.

TABLE 1. Evaluation of startup procedures.

Model	Evaluation		
	VQA	ZS IG. (↓)	ZS IN.
Init. Text Unicoder from CoCa	68.45	11.41	75.34
Train from Scratch	68.34	11.58	74.89

This investigation builds upon foundational studies [24], [28] that examine the integration of visual and textual data within a multimodal framework, particularly focusing on the Visual Question Answering (VQA) task as an experimental paradigm. In this context, the VQA task is reconceptualized as a classification problem, targeting the 3,129 most frequently provided answers in the dataset. The methodology involves introducing raw images to an image unicoder in encoding mode while simultaneously processing questions through a text unicoder in decoding mode. Following this,

FIGURE 7. Subjective outcomes of zero-shot image synthesis using ZsU-IT-large.

a Bimodal decoder is activated to integrate the outputs from the text decoder and correlate them with the encoded visual data. This integration results in a unified global feature vector derived from the Bimodal decoder's final token, which forms the basis for predicting answers through a linear classifier trained on this composite feature. Comparative analysis, as shown in Figure 3, demonstrates the robust performance of our approach relative to existing Vision-Language Pre-training (VLP) models.

In the domain of Stanford Natural Language Inference-Visual Entailment (SNLI-VE), our model surpasses existing VLP frameworks in accurately predicting the relational dynamics between entities. The relational judgments are primarily supported by the comprehensive output feature of the Bimodal decoder, which is processed through a linear classifier, with detailed results presented in Figure 3. It is noteworthy that while standard models, including our own, typically rely on visual inputs as premises, the OFA model distinctively combines both visual and textual data.

Further, this research explores the domain of text-to-image synthesis, specifically through fine-tuning on the MS-COCO training dataset following methodologies outlined in [18] and [29]. The efficacy of our model in creating photorealistic images from textual descriptions is evaluated using the Fréchet Inception Distance (FID) metric, across a subset of 30,000 test images. This metric helps quantify the model's ability to generate high-quality images that closely resemble real-life visuals, highlighting its potential in realistic image synthesis from diverse text prompts.

A comprehensive series of ablation studies have been conducted to illuminate the architectural and procedural subtleties of our model, primarily employing its base configuration for these evaluations. Modifications were made to batch sizes and training durations, with a total batch size distributed as 4,352callocated between contrastive/Image-to-Text (I2T) loss (4,096) and Text-to-Image (T2I) loss (256). The model underwent a streamlined training protocol of 200,000 steps, intentionally omitting high-resolution pre-training to assess baseline capabilities. Performance assessments covered a spectrum of critical tasks including zero-shot ImageNet Classification, fine-tuned Visual Question Answering (VQA), and MS-COCO zero-shot Captioning (evaluated using CIDEr scores), with lower scores indicating superior image quality). These ablation results underscore the model's adaptability and effectiveness across a varied set of vision-language tasks, demonstrating its potential to significantly propel advancements in the field.

In this research, we delve into a meticulous ablation study concentrating on three key training objectives: contrastive loss, Image-to-Text (I2T) loss, and Text-to-Image (T2I) loss. The study aims to explore their synergies and potential conflicts. Initial observations from contrasting the results at the extremities of the dataset indicate that the integration of bi-modal generative objectives modestly enhances image comprehension capabilities, as evidenced by a 0.3% improvement in zero-shot ImageNet accuracy, surpassing what is achievable through contrastive loss alone. Further analysis of intermediate data reveals a notable contradiction between the generative losses: the inclusion of T2I loss results in a decrease of 2.6 points in VQA performance, whereas the addition of I2T loss leads to a 0.6 point increment in the zero-shot image generation FID score, suggesting a delicate balance between these objectives. An exhaustive ablation process was employed to determine the optimal weighting of loss coefficients for these objectives, which was then implemented across all experimental setups. Comprehensive documentation of this process and the derived conclusions are available in the supplementary materials accompanying this study.

These ablation studies provide critical insights into the interplay of different loss functions within our model and highlight the importance of fine-tuning training parameters to optimize performance across diverse vision-language tasks. The results not only affirm the model's robustness and versatility but also pave the way for further enhancements in multimodal learning frameworks.

The exploration of Vision-Language integration techniques has predominantly adhered to an encoder-decoder framework, where distinct encoders process image and text inputs, subsequently merged and interpreted by a Bi-modal decoder. Diverging from this convention, our study introduces the concept of a unicoder, a versatile entity capable of both encoding and decoding within a single modality, employing a unified set of parameters for both functions. This novel approach contrasts traditional encoders by facilitating parameter-efficient unimodal representation processing without inflating the model's parameter count. Supplementary materials offer a visual comparison between the traditional encoder-focused models and our unicoder paradigm.

Additionally, our research probes into the feasibility of commencing the training process from an initial state, devoid of any pre-trained weights. Specifically, although the text unicoder was initially equipped with weights derived from a pre-trained unimodal text decoder from CoCa, we explored an alternative approach where the entire model underwent training from scratch. This experiment aimed to evaluate the necessity and impact of using pre-trained weights by setting up a control condition that lacked such initial advantages. The comparison was rigorously controlled, with standardized batch sizes employed to ensure uniform training conditions across the two setups, and the results are collated in Table 1.

The experimental findings suggest that the benefits of pre-loading CoCa's pre-trained weights are minimal, enhancing performance by only 0.3% and 0.2, respectively. Additionally, this strategy showed no significant improvement in the Visual Question Answering (VQA) performance, underlining the marginal role pre-trained weights play in this context. These results support the feasibility of initiating the ZsU-IT model's training from scratch, effectively challenging the established presumption that pre-existing weights are critical for achieving robust performance in Vision-Language tasks. This insight highlights the ZsU-IT model's inherent capabilities and suggests a potential reduction in dependency on pre-trained models, paving the way for more flexible and foundational approaches in training Vision-Language models.

D. GRAPHICAL REPRESENTATION

In this phase of our investigation, the Universal Image Transformer (ZsU-IT) not only showcased outstanding performance across numerous zero-shot tasks but also its capability to generate high-quality images from textual descriptions. This proficiency was rigorously analyzed through visualizations that portrayed both the achievements and obstacles faced by the ZsU-IT-Large model in image generation tasks, as demonstrated in Figure 7. These visual examples affirm ZsU-IT's skill in producing complex, open-domain imagery based on textual prompts. Particularly, the successful instances highlight the model's nuanced ability to interpret and visually represent intricate details such as textures, forms, and colors as described by the text inputs.

Conversely, the less successful outputs provide critical insights, revealing the current limitations of the model and identifying potential areas for improvement. These cases illustrate the sophisticated relationship between textual inputs and visual outputs, emphasizing the inherent challenges in achieving consistent and accurate image synthesis across varied and sometimes vague prompts. The findings from these visual assessments demonstrate the effectiveness of our pre-training approach and confirm the ZsU-IT's ability to support a wide range of applications across different domains. This study represents a significant progression in the fields of visual representation learning and text-to-image generation, setting a promising course for continued research and development in this vibrant and expanding area.

V. CONCLUSION

This study has introduced ZsU-IT, a cutting-edge vision-language foundation model that integrates three essential objectives. The architecture of ZsU-IT is structured around three core components: an image encoding unicoder, a textual encoding unicoder, and a cross-modal attention mechanism for decoding. These modules are ingeniously crafted to switch between dual operational modes, supporting both uni-modal and cross-modal encoding and decoding functions. The ZsU-IT framework has undergone rigorous training on a large dataset comprised of web-crawled image-text pairs and carefully annotated image datasets. Our empirical findings demonstrate that ZsU-IT excels in zero-shot learning and transfer capabilities across a diverse range of tasks that are crucial for visual comprehension, including uni-modal visual analysis, the alignment of images with their textual descriptions, and a comprehensive understanding of image-text interactions. Additionally, ZsU-IT has proven to be highly effective in producing high-quality, diverse images from textual descriptions, highlighting its potential in tasks involving creative content generation. Consequently, the ZsU-IT model not only enhances our understanding of the complex dynamics between visual and textual data but also opens up new avenues for research and application in this rapidly evolving field.

ACKNOWLEDGMENT

(Yujia Gu, Brian Liu, and Tianlong Zhang contributed equally to this work.)

REFERENCES

[1] J. B. Alayrac, "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23716–23736.
[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.
[3] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, Y. Li, and D. Krishnan, "Muse: Text-to-image generation via masked generative transformers," 2023, *arXiv:2301.00704*.
[4] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "MaskGIT: Masked generative image transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11305–11315.
[5] M. Ding, W. Zheng, W. Hong, and J. Tang, "CogView2: Faster and better text-to-image generation via hierarchical transformers," 2022, *arXiv:2204.14217*.
[6] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, "Make-a-scene: Scene-based text-to-image generation with human priors," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv-Yafo, Israel, 2022, pp. 89–106.
[7] J. Ho and T. Salimans, "Classifier-free diffusion guidance," 2022, *arXiv:2207.12598*.
[8] C. Jia, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.
[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021.
[10] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2200–2207, doi: 10.1109/ICCV.2013.274.
[11] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-IO: A unified model for vision, language, and multi-modal tasks," 2022, *arXiv:2206.08916*.
[12] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision meets language-image pre-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 529–544.
[13] H. Pham, Z. Dai, G. Ghiasi, K. Kawaguchi, H. Liu, A. W. Yu, J. Yu, Y. T. Chen, M. T. Luong, and Y. Wu, "Combined scaling for open-vocabulary image classification," 2021, *arXiv:2111.10050*.
[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[16] A. Ramesh, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.
[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 10674–10685, doi: 10.1109/CVPR52688.2022.01042.

[18] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. G. Lopes, B. K. Ayan, and T. Salimans, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.
[19] M. Tschannen, B. Mustafa, and N. Houlsby, "CLIPPO: Image-and-language understanding from pixels only," 2022, *arXiv:2212.08045*.
[20] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "GIT: A generative image-to-text transformer for vision and language," 2022, *arXiv:2205.14100*.
[21] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," 2022, *arXiv:2202.03052*.
[22] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: BEiT pretraining for all vision and vision-language tasks," 2022, *arXiv:2208.10442*.
[23] Y. Wang, J. Chen, M. Wang, H. Li, W. Wang, H. Su, Z. Lai, W. Wang, and Z. Chen, "A closer look at classifier in adversarial domain generalization," in *Proc. 31st ACM Int. Conf. Multimedia*, vol. 31, Oct. 2023, pp. 280–289.
[24] Z. Wang, J. Yu, A. Wei Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," 2021, *arXiv:2108.10904*.
[25] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "FILIP: Fine-grained interactive language-image pre-training," 2021, *arXiv:2111.07783*.
[26] H. You, L. Zhou, B. Xiao, N. Codella, Y. Cheng, R. Xu, S. F. Chang, and L. Yuan, "Learning visual representation from modality-shared contrastive language-image pre-training," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 69–87.
[27] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved VQGAN," 2021, *arXiv:2110.04627*.
[28] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.
[29] J. Yu, Y. Xu, J. Yu Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. Karagol Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, "Scaling autoregressive models for content-rich text-to-image generation," 2022, *arXiv:2206.10789*.
[30] L. Yuan, "Florence: A new foundation model for computer vision," 2021, *arXiv:2111.11432*.
[31] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "LiT: Zero-shot transfer with locked-image text tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18102–18112.
[32] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 13041–13049.

YUJIA GU received the M.F.A. degree in graphic design, with a focus on information experience design and cross-culture design area. In her studies, she is currently working on conceptualizing, designing, and implementing multimedia, and critical thinking, with a focus on cultural phenomena. In the past six years, she was with China Construction Shenzhen Southern Company and was specifically responsible for the delivery of advertising, media, and property aspects in project operation. She has been working on visual communication and media design, since 2015. She is collaborating with China Railway Rolling Stock Corporation as a Special Consultant and leads the design ideation for CRRC's distinctive train design and developing signage and wayfinding systems. She is also a Senior Partner and the Director of digital media of Henan Yujia Brand Design Company Ltd., which complemented her design methodology with practical insight from the field. She is collaborating with multiple art and design companies, such as SIA International Education, Absorb, and AndLab Studios.

BRIAN LIU is currently pursuing the degree with the Stuyvesant High School, New York, NY, USA. He was a Scholarship Winner and an USACO Silver Qualifier. He has a strong passion for computer science and related applications, including data science, AI, and modeling. He volunteers for his school's journal taking charge of web development and content management. Outside school, he attended a few STEM boot camps, including MIT APP Inventor Camp and All-Star Code Intensive Camp, where he explored a variety of technology subfields, including app development, AI, and machine learning. Last summer, he had a work experience program in partnership with Amazon Web Services, where he developed a few machine learning models for AWS. He also did an internship with a tech startup to research sentence analysis functions and NLP-powered tools to enhance writing. He has been participating in Kaggle competitions to deepen his understanding of how modeling and AI techniques can be applied to solve real problems.

TIANLONG ZHANG is currently pursuing the master's degree in information science with the University of Pittsburgh, with a focus on big data analytics. He is enthusiastic about extracting and analyzing large datasets to find important insights, particularly in using these methods to address real-world challenges.

XINYE SHA received the B.A. degree in mathematics from the Homerton College, University of Cambridge, and the M.A. degree in mathematics of finance from Columbia University, New York, NY, USA. He is currently a Quantitative Developer in the financial industry. He is interested in artificial intelligence and its applications in finance, especially deep learning in finance.

SHIYONG CHEN received the B.S. degree in communication engineering from Beijing Jiaotong University (BJTU), Beijing, China, in 2021. He is currently pursuing the Ph.D. degree in communication and information systems with the School of Electronics and Information Engineering. His current research interests include machine learning and AI.