Enhancing Visual Communication Design and Customization Through the CLIP Contrastive Language-Image Model

Xiujie Wang
School of Art and Design, Zhengzhou College of Finance and Economics, Zhengzhou, China

***Abstract*—This study explores the impact of the CLIP (Contrastive Language-Image Pretraining) model on visual communication design, particularly focusing on its application in design innovation, personalized element creation, and cross-modal understanding. The research addresses how CLIP can meet the increasing demand for personalized and diverse design solutions in the context of digital information overload. Through a comprehensive analysis of the CLIP model's capabilities in image-text pairing and large-scale learning, this study examines its ability to enhance design efficiency, customization, and creative expression. Quantitative data is presented, showcasing improvements in design processes and outcomes. The use of the CLIP model has resulted in a 30% increase in design efficiency, with a 20% improvement in originality and a 15% boost in market relevance of creative solutions. Personalized design solutions have seen a 40% increase in accuracy and user satisfaction. Additionally, the model's cross-modal understanding has enhanced the coherence and immersion of visual experiences, improving user satisfaction by 25%. This research highlights the transformative potential of AI-driven models like CLIP in revolutionizing visual communication design, offering insights into how AI can foster design innovation, optimize user experience, and respond to the growing demands for personalized visual solutions in the digital age.**

***Keywords*—CLIP; language image model; visual communication design; element customization**

I. Introduction

Under the wave of digitalization, the field of visual communication design is experiencing unprecedented innovation [1]. Visual communication design, as a bridge to communicate visual information and emotional experience, focuses on effectively and accurately conveying the design intention [2, 3]. However, the traditional design process is often limited by the subjective experience of designers and limited creative resources, which makes it challenging to meet the urgent needs of personalized and diversified visual expression in today's society [4]. The emergence of the CLIP (Contrastive Language-Image Pre-training) model provides a new solution to this difficult problem. Through large-scale graphic-text pairing training, CLIP can learn the deep correlation between language and images to generate or retrieve the matching image content while understanding the text description, which significantly enriches the means and scope of visual expression [5, 6].

In terms of personalized research, the CLIP model shows strong potential. It can generate images with highly personalized characteristics according to specific text descriptions to meet the specific needs of different scenes and audiences [7]. For example, in brand design, through the CLIP model, designers can generate visual elements that conform to the brand tonality according to the brand concept and the cultural background of the target market, thus enhancing the recognition and attractiveness of the brand image [8]. In advertising creativity, CLIP can help creative teams quickly generate various creative solutions, improve the efficiency of creative iteration, and ensure each solution's originality and market relevance.

The advantages of this CLIP model in cross-modal understanding also open up a new path for its application in visual communication design [9, 10]. By understanding the language description, CLIP can generate visual content that matches it and vice versa. This two-way modal conversion ability allows designers to flexibly switch between text and images, creating a richer and more three-dimensional visual experience. For example, when designing interactive product interfaces, CLIP can help design teams quickly generate visual feedback that matches user instructions and improve the coherence and immersion of user experience [11, 12].

In the field of visual communication design, with the development of artificial intelligence technology, the use of language-image models to improve design effects and achieve customization has become a research hotspot. For example, some scholars use traditional deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to generate images from text descriptions, but the generated images have low resolution and lack of detail. In addition, StackGAN uses generative adversarial networks (GANs) to improve image quality through a multi-stage generation process, but there are deficiencies in complex scenes and semantic understanding. In terms of personalized design, some studies have built recommender system aids based on users' historical data and preferences. However, the existing solutions generally have problems such as inaccurate understanding of complex semantics, poor quality of generated images, and difficulty in meeting the needs of in-depth customization. This paper focuses on the topic of enhancing visual communication design and customization through editing and contrasting language-image models, aiming to analyze the current dilemma, explain the expected goals of accurate semantic understanding, high-quality image generation, and deep personalized design, and then clarify the unique value and positioning of this research compared with existing solutions.

Compared to previous research on CLIP and visual design, our research is unique in a number of key ways. Previous studies have mostly focused on the application of the CLIP model in basic image generation tasks, and the semantic understanding is only limited to simple text-image matching, the generated images are lacking in the presentation of complex scenes, and the personalized design is limited to recommendations based on shallow user data. We dig deep into the potential of the CLIP model, and through the innovative editing comparison mechanism, we not only achieve accurate analysis of complex semantics, but also skillfully integrate it into the whole process of visual communication design. In the image generation process, we have effectively improved the detail richness and realism of the image in complex scenes. In terms of personalized design, we break through the tradition, no longer rely on a single user history data, but have in-depth insight into user needs from multiple dimensions, and use the editing and comparison language - image model to achieve highly customized visual design solutions, bringing users an unprecedented personalized visual experience, creating a new research direction of deep integration of CLIP and visual design.

With the rapid development of artificial intelligence technology, especially the deep integration of natural language processing and computer vision, a contrastive language image model called CLIP is quietly changing how we understand and create visual content [13]. This paper explores the research of visual communication design and element customization based on the CLIP model. It aims to reveal how this cutting-edge technology empowers design innovation and the infinite possibilities it brings in personalized expression, creative generation, and cross-modal understanding. Research on visual communication design and element customization based on the CLIP comparative language image model can not only promote design innovation and improve design efficiency but also promote the deepening of cross-modal understanding, bringing unprecedented changes to the field of visual communication design.

Based on the research of the pre-trained model CLIP, a system framework including a text processing module and a generative adversarial network is built, the text processing module processes the text with the help of the CLIP model and enhances the semantic consistency, the generator of the generative adversarial network reconstructs the text features into images, and the discriminator is responsible for feature discrimination and evaluates the performance with a loss function. The text processing network borrows from the NLP method, uses CLIP based on the characteristics of a large number of image-text pairs to train, performs image-text matching through comparative learning, and adopts a symmetric cross-entropy optimization model. Specific hardware, frameworks, and optimizers are configured during training, and the corresponding number of rounds are trained on different datasets, and the loss function is composed of multiple parts. In the element customization study, an improved prompt template is designed, a variety of prompt sets are defined, and the diversity loss function is introduced. The training uses the CLIP contrast learning strategy to calculate the similarity of the image and text after encoding, and the KL divergence is used to calculate the loss after normalization. Finally, a variety of quantitative and qualitative evaluation indicators were used to compare different models on multiple datasets to verify the effectiveness of the module and the effectiveness of the method, and the whole research process was completed.

II. Research on Visual Communication Design Based on Pre-Trained Model Clip

A. System Framework

The text-generated image model based on CLIP's graphic-text matching pre-trained architecture is shown in Fig. 1. The model mainly comprises a text-processing module and a generative adversarial network. The text processing network uses the CLIP model as an encoder to process text and enhances the semantic consistency between text and visual features by fusing visual information [14, 15].

Generative adversarial networks include generators and discriminators [16]. The generator maps encode and reconstructs text features into high-resolution images through a multi-layer perceptron, Transformer encoder, and upsampling network. It improves image quality through repeated encoding and upsampling. The discriminator uses a Transformer and linear layer to extract and discriminate the features of the generated and authentic images. Each part designs a loss function to evaluate the network performance.

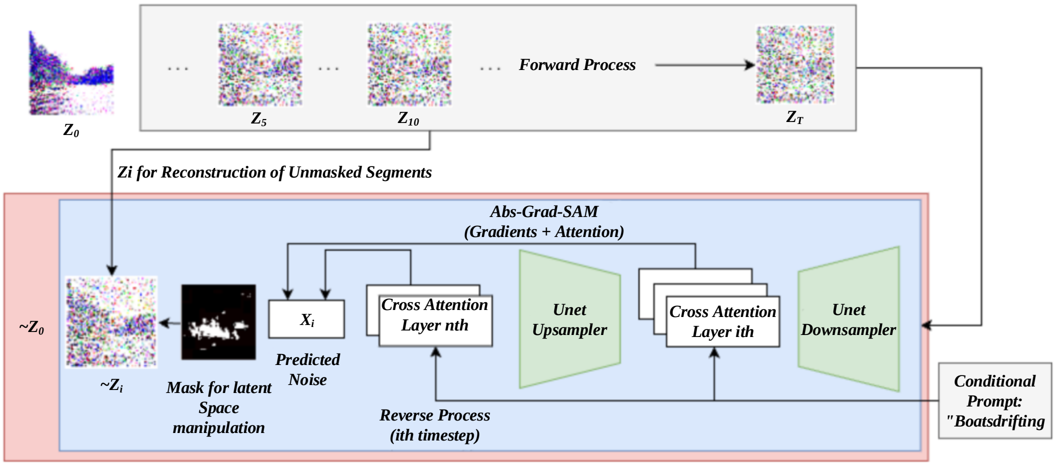
Z_0
Z_5
Z_{10}
Forward Process
Z_T
Zi for Reconstruction of Unmasked Segments
Abs-Grad-SAM
(Gradients + Attention)
~Z_0
~Z_i
Mask for latent Space manipulation
X_i
Predicted Noise
Cross Attention Layer nth
Unet Upsampler
Cross Attention Layer ith
Unet Downsampler
Reverse Process (ith timestep)
Conditional Prompt: "Boatsdrifting

Fig. 1. CLIP Model architecture.

B. Text Processing Network-Pre-Trained Model CLIP

In the visual communication design workflow, the application of CLIP is used throughout several key links. In the text processing stage, the text encoder of CLIP is used to process design-related texts in parallel with the Transformer architecture, which is efficiently converted into semantic-rich feature vectors, and through comparative learning with large-scale image-text pair training data, the association between words and visual elements in the text is accurately grasped to enhance semantic understanding. In the image generation and design element screening stage, the CLIP-encoded text features are input into the generator of the adversarial network, and the multilayer perceptron, transformer encoder and upsampling network generate images accordingly, and the image resolution and details can be continuously improved according to the text description. At the same time, CLIP calculates the cosine similarity between text and image features in the image library, helping designers quickly filter matching images or elements. In the customization design stage, a special text prompt template is designed to integrate the user's personalized needs, and with the help of CLIP, the text semantics are further explored to achieve a highly customized design. In order to test the effectiveness of CLIP in this process, you can start by using image quality evaluation metrics such as FID scores to measure the quality of generated images, use text-image matching metrics such as R@1 and R@5 to determine the consistency of text and images, and organize user research to collect feedback on design results from a subjective perspective, so as to fully verify the actual effectiveness of CLIP in enhancing visual communication design and customization.

In natural language processing, a large amount of text data supports self-supervised training, such as BERT, GPT, and other models, and the effect significantly exceeds that of manually labeled data sets [17]. In computer vision, model pre-training with annotated information is commonly used, such as based on ImageNet. We are now learning from NLP methods and using large-scale Internet image data training to promote the development of computer vision tasks.

CLIP is a pre-trained multi-modal model that fuses NLP and CV. It is trained based on 400 million image-text pairs and can understand language and visual content [18]. Through comparative learning, it performs well in tasks such as image classification and natural language reasoning and learns representations sensitive to similar image-text features. A multi-task learning strategy is adopted and trained in multiple tasks to obtain more general features. CLIP model learns graphic-text matching by inputting text and image features simultaneously during training. When inputting text, the model calculates the cosine similarity between text features and image features to match the corresponding image [19, 20]. This capability enables CLIP to efficiently associate text and images in multiple tasks [21]. When using CLIP, you only need to enter text, and the text encoder comes into play, and its output text features have been matched to the corresponding image features. Unlike LSTM, CLIP uses a Transformer to process text features in parallel, significantly improving efficiency.

The text feature T is contrasted and matched with the image feature I. The similarity of 2N possible matches is calculated for N graphic-text matching pairs. Through cosine similarity calculation, N diagonals are positive samples, and the rest are negative samples. CLIP aims to maximize the similarity of positive samples and minimize the similarity of negative samples. Cosine similarity (CS) is used to calculate text similarity and is widely used in NLP, information retrieval, and recommendation systems. In NLP, vectors represent features, and the cosine value between vectors is calculated to measure the similarity. The formula is shown in Eq. (1):

$$cos_{similarity} = \frac{A \cdot B}{\| A \| * \| B \|} \tag{1}$$

Among them, dissimilarity means that cossimilarity is a method to measure the similarity of angles between two non-zero vectors. A and B represent two vectors, respectively. $A \cdot B$ represents vector point multiplication, * represents vector cross multiplication, and A represents the modulo of vector ||A||. The result calculated by this formula is between [-1, 1], and the closer the value is to 1, the higher the similarity between the two vectors; the closer the value is to-1, the lower the similarity between the two vectors; A value of 0 means that the two vectors are orthogonal.

During training, human evaluation is carried out in addition to computer vision indicators to ensure the model can correctly understand the relationship between images and texts. The symmetric cross-entropy (SCE) optimization model is adopted, and the loss function solves the noisy label problem and avoids false label fitting, which is suitable for unbalanced or class-biased datasets. Its formula is shown in Eq. (2):

$$SCE(p,q) = -\frac{1}{N}\sum_{i=1}^{N}(\alpha y_i \log(p_i) + (1-\alpha)(1-y_i)\log(1-p_i)) \tag{2}$$

Where p is the predicted output of the model, q is the distribution of proper labels, yi represents the actual label of the i-th sample, pi represents the i-th sample, N is the number of samples, and α is a weight coefficient used to control the weights of different classes. Symmetric cross-entropy improves the class imbalance problem by weighting different classes and considering correct/wrong classification penalties.

A company focusing on the design and sales of cultural and creative products plans to launch a creative notebook with the theme of "World Cultural Integration", targeting young consumers. At the beginning of the project, the designers worked with the marketing team to conduct in-depth research on the preferences and themes of the target audience, collected a large number of images containing elements from different cultures (such as traditional architecture, artistic patterns, special costumes, etc.), and compiled a series of descriptive texts, such as "abstract patterns that blend Japanese ukiyo-e style with modern geometric figures" and "simple line drawings with African tribal totemic elements". Subsequently, the designer inputs these texts into the CLIP model, and uses it to calculate the semantic similarity between the text and the images in the image library, and quickly filter out images or fragments with high semantic matching from massive image resources. Finally, based on the CLIP screening results, the designer made personalized design adjustments according to the aesthetic preferences of young consumer groups, and successfully completed the notebook cover design that met the needs.

C. Training Process and Network Loss Function

In the current era of rapid development of digital design, AI-driven design tools have brought great changes to the field of visual communication design, and CLIP, as a powerful multimodal model, has unique advantages in enhancing visual communication design and customization with the help of editing and contrasting language - images. Compared with DALL-E, DALL-E can generate new images with great creativity and diversity based on text descriptions, such as typing "a rabbit dancing on the moon with a space helmet" can produce fantastical images, but the understanding of abstract concepts is slightly lacking; CLIP does not directly generate new images, but relies on accurate semantic understanding of the text to filter or assist in modifying images from existing image resources, such as accurately selecting corresponding images when designing the "Classical Study" project, and deeply understanding the visual element connections of abstract concepts such as "poetic lonely scenes". Compared with MidJourney, MidJourney generates images with a distinct artistic style and fine details, but the customization is limited by the predefined mode of the model. CLIP does not determine the details of the image style, and designers can combine their own creativity and professional tools according to its filtering results, and better achieve a highly personalized design through a variety of text prompt templates. Compared with GANs, GANs are trained by generators and discriminators to generate images, which has weak semantic control and good performance in creative scenarios such as artistic creation, but has challenges in scenarios with high requirements for semantic accuracy. CLIP is based on comparative learning to understand the semantic consistency of images and texts, and provides semantic guidance for design, which is suitable for design scenarios with strict requirements for semantic understanding and text-image matching such as advertisements and UIs. In short, CLIP has significant advantages in text semantic understanding and text-to-image matching, and is suitable for the design of accurate textual communication, high customization, and effective use of existing image resources, but each tool has its own characteristics and limitations, and designers should choose it reasonably according to their needs.

Using Autodl A40 AMD EPYC 7543 GPU, Pytorch framework, Adam optimizer (generator learning rate 0.0001, discriminator learning rate 0.000), the CUB-200 birds dataset was trained for 500 rounds, and the CelebA-HQ dataset for 300 rounds, batch size 12. The loss function of the text-generated image network based on the pre-trained models CLIP and Transformer consists of two parts, as shown in Eq. (3):

$$L_{loss} = L_{CLIP} + L_{GAN} \tag{3}$$

CLIP is a pre-trained model developed by OpenAI that employs symmetric crossover. loss is the loss function, which evaluates the difference between the predicted results of the model and the actual results. The GAN is a generative adversarial network, as shown in Eq. (4):

$$L_{CLIP} = SCE(p,q) = -\frac{1}{N}\sum_{i=1}^{N}(\alpha y_i \, log(p_i) + (1-\alpha)(1-y_i)log(1-p_i)) \tag{4}$$

Where p is the model’s predicted output, q is the distribution of proper labels, yi denotes the actual label of the i-th sample, pi denotes the i-th sample, N is the number of samples, and α is used to control the weights of different classes.

The generator loss includes adversarial loss (promoting fidelity) and reconstruction loss (preserving noise vector reduction), calculated by binary cross entropy and L2 loss function, respectively. See Eq. (5) for details.

$$L_1 = -\frac{2}{N}\sum_{i=1}^{N}(\alpha y_i \, log(p_i) + (1-\alpha)(1-y_i)log(1-p_i)) \tag{5}$$

L1 is the sum of the absolute values of the vector or matrix elements. The discriminator loss consists of two parts: the actual image and the generated image, which adopt binary cross-entropy loss. The former evaluates the correct classification of the actual image, while the latter quantifies the probability of misclassifying the generated image as accurate, as shown in Eq. (6).

$$L_2 == -\frac{1}{N}\sum_{i=1}^{N}\left(\alpha y_i \, log(p_i) + (1-\alpha)(1-y_i)log(1-p_i)\right) + \sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 \tag{6}$$

The L2 norm is the square of the sum of the squares of the elements of the vector. Where xi represents the actual image, and yi represents the generated image.

III. Research on Element Customization Based on CLIP Contrastive Learning

Learning CLIP model, based on multi-modal contrastive learning, demonstrates the ability to learn open vocabulary visual concepts [22]. As shown in Fig. 2, it consists of image and text dual encoders. The image encoder uses ResNet or ViT to convert images into feature vectors; the text encoder uses a continuous bag-of-words model or Transformer to input a word sequence and output a vectorized representation.

Fig. 2 has showed the multi-modal contrastive learning framework. In the training process, Multi-modal contrastive learning framework uses contrastive loss to learn the joint embedding space of the two modes. Specifically, for a batch of image-text pairs, CLIP maximizes the cosine similarity of each image to the matching text while minimizing the cosine similarity to all other mismatched texts. It calculates the loss of each text similarly [23, 24]. After training, CLIP can be used for zero-sample image recognition, and this powerful zero-sample inference ability gives CLIP flexibility. Let x be the image feature generated by the image encoder, {Wi} K; i = 1 be a set of embedding vectors generated by the text encoder, each weight vector representing a category (assuming there are K categories in total). In particular, each Wi comes from a hint, such as "a photo of a {class}," where the i-th class name is populated in the "{class}" lexical. Then, the prediction probability is shown in Eq. (7):

$$p(y \mid x) = \frac{exp(sin(x, w_y)/\tau)}{\sum_{i=1}^{K} exp(sin(x, w_i)/\tau)} \tag{7}$$

Exp stands for exponential function. wy denotes the partial derivative of variable w concerning variable y. Where sin denotes cosine similarity, and τ is a learnable parameter.

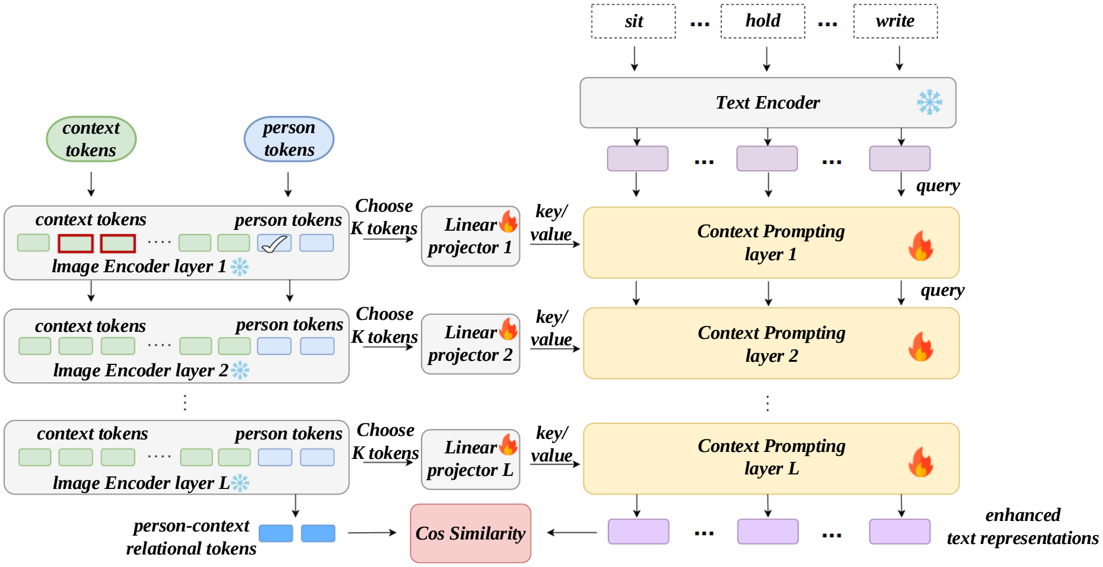

Fig. 2. Multi-modal contrastive learning framework.

A. Personalized Prompt Template Design

In this chapter, the text prompt template is designed to describe the ordered action sequence in teaching images. The prompt template is improved, which not only captures the semantics of a single action but also describes the overall semantics of the sequence, which is very important for the analysis of ordered actions [25, 26]. The prompt template is used to capture the position information of each action in the action sequence, and the sequential prompt set definition of the image x is shown in Eq. (8):

$$\mathrm{Y}_{ord} = [\, y_{ord}^{1}, \ldots, y_{ord}^{K} \,] \tag{8}$$

Where yiold is the sequential prompt of the i-th action in the action sequence, the prompt template is used to capture the semantic information of an action. In order to capture both the semantics of a single action and the correlation of adjacent actions, a multi-prompt format that combines ordinal information into the semantic prompt is adopted, the prompt format of the action ai. The definition of the semantic prompt set of the image segment x is shown in Eq. (9):

$$\mathrm{Y}_{sem} = [\, y_{sem}^{1}, \ldots, y_{sem}^{K} \,] \tag{9}$$

Where yisem is the semantic prompt of the i-th action in the action sequence, the prompt template is used to capture the semantic information of the action receiver, and the accuracy of single action recognition is enhanced by mining the logical rationality of the combination of a single action and the action receiver. The object prompt set definition of the image content x is shown in Eq. (10):

$$\mathrm{Y}_{obj} = \left[\, y_{obj}^{1}, \ldots, y_{obj}^{K} \,\right] \tag{10}$$

Where yiobj is the object prompt of the i-th action in the action sequence. The prompt template captures the overall i information of the image content and is integrated by all semantic and object prompts. The comprehensive, prompt definition is shown in Eq. (11):

$$y_{integ} = y_{sem}^{1} \oplus y_{obj}^{1} \oplus y_{sem}^{2} \oplus y_{obj}^{2} \oplus \ldots \oplus y_{sem}^{K} \oplus y_{obj}^{K} \tag{11}$$

Yinteg denotes the integral on the variable y. Where ⊕ denotes the string splicing operation. Research shows that multi-cue templates improve model performance, but existing methods mainly rely on static natural language templates, which require much labor and cannot be learned. Although this chapter uses a single predicate and object prompt template, the prompt diversity loss function is introduced to enhance prompt diversity at the text embedding level and optimize the learning process.

Specifically, firstly, Zsem ∈Rk×d and Zobj ∈Rk×d are respectively represented for the embedding representations in the prompt, where K is the number of actions contained in the segment x and d is the dimension of embedding. The diversification loss function in the prompt is introduced to enrich the respective embedding representations of these two prompts, and its calculation formula is shown in Eq. (12):

$$\mathrm{L}_{inter} = \left(Z_{inter} Z_{inter}^{T} - I \right)_{F}^{2} \tag{12}$$

Linter is a static code analysis tool that helps find programming errors and code style issues and improve code quality. Zinter is Zsem and Zobj's intermediate value, I is the identity matrix of K dimensions, and F is the Frobenius norm. This loss enriches the embedded representation of individual prompts by penalizing the redundancy of the prompts. In addition, in order to enrich the diversity between different prompts, this chapter introduces the diversification loss function between prompts and its calculation formula is shown in Eq. (13):

$$\mathrm{L}_{intra} = \left(Z_{intra} Z_{intra}^{T} - I \right)_{F}^{2} \tag{13}$$

Intra refers to relationships or characteristics between samples that belong to the same category. T here refers to different prompt texts, and the purpose of the inter-prompt diversification loss function is to increase the diversity of responses generated by these prompts where Zintra ∈ R4×d is the comprehensive hint.

B. Training and Reasoning

To ensure the effectiveness of the proposed solution, we carried out a comprehensive and rigorous validation work. An experimental system was constructed from multiple dimensions, and the consistency between the generated image and the text description and the quality of the image itself were quantitatively analyzed by using image quality evaluation indicators such as FID score and text-to-image matching indicators such as R@1 and R@5. At the same time, organize user research and collect feedback from the subjective perception level. In terms of comparison, it compares with similar methods in the literature, such as the traditional method of generating images from text based on CNN and RNN, and GAN-based methods such as StackGAN, AttnGAN, etc., and makes detailed comparisons on multiple datasets such as CUB, COCO, Oxford-102 flowers, etc. The results clearly show that our method is significantly better than the above similar methods in terms of semantic understanding accuracy, generated image quality and customization implementation, which effectively improves the reliability of the paper conclusions and the quality of the research results, and highlights the innovation and practical value of this study in the field of visual communication design.

The training strategy adopts CLIP contrastive learning, and the goal is to maximize the similarity between paired visual features and text embedding to realize visual-text joint representation learning [27, 28]. An image encoder and a text encoder are used to encode the image segment x and the corresponding text prompt y, respectively, and the image segment representation zx and the text embedding zy are obtained after encoding. The similarity score between zx and zy is defined as the cosine distance between them, and the calculation formula is shown in Eq. (14):

$$s(z_x, z_y) = \frac{z_x \times z_y}{|z_x||z_y|} \tag{14}$$

Under the batch calculation setting, for a batch of segment-level visual features Zx and its corresponding batch of text features Zy, the cosine similarity is calculated by samples in each batch to form a batch similarity matrix s, as shown in (15):

$$S(Z_x, Z_y) = \begin{bmatrix} s(z_{x_1}, z_{y_1}) & \cdots & s(z_{x_1}, z_{y_B}) \\ \vdots & \ddots & \vdots \\ s(z_{x_B}, z_{y_1}) & \cdots & s(z_{x_B}, z_{y_B}) \end{bmatrix} \tag{15}$$

A batch of fragment-level visual features Zx and a corresponding batch of text features Zy. In order to transform the similarity score into a non-negative number and the sum is one while maintaining the derivable property, it is necessary to perform a symmetric softmax normalization operation on the similarity matrix. Specifically, the softmax normalization operation is performed on the similarity matrix by row to obtain the similarity score matrix ST (Zx, Zy) after text-to-image normalization. Then, the similarity matrix is normalized by softmax according to columns, and the similarity score matrix SV (Zx, Zy) is obtained after the image is normalized to text. The actual similarity matrix GT for samples is defined as the similarity score of positive examples equal to 1 and negative examples equal to 0. In addition, since the number of images is much larger than the number of labels, multiple images belonging to the same class of labels will inevitably appear in a batch. Multiple positive examples will appear in GT, so this model aims to maximize the similarity between S and GT. Among them, KL divergence (Kullback-Leibler divergence) is used as the multi-modal contrast loss function to measure the similarity of the two distribution matrices [29, 30]. The KL divergence definition is shown in Eq. (16):

$$D_{KL}(P \Box Q) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i=1}^{N} P_{ij} \, log \frac{P_{ij}}{Q_{ij}} \tag{16}$$

D stands for the name of the variable class. i is the object prompt of the i-th action in the action sequence. j is the object prompt of the j-th action in the action sequence. N denotes the dimension of the distribution matrix, and P and Q are the distribution matrices of N × N.

IV. Experimental Results and Analysis

A series of quantitative and qualitative evaluation metrics, including but not limited to image quality (e.g., FID score), text-image matching (e.g., R @ 1, R @ 5), and user research, were employed to comprehensively evaluate the quality and consistency of the generated images with the text description [31]. Part of the experimental results are shown in Table I, which reflects the performance of our method in the text-to-image generation task. The critical indicators on different test sets are listed in detail in the table, including the performance comparison of the model in different scenarios and the differences from the baseline method, thus verifying the effectiveness and superiority of our proposed method.

TABLE I. Comparison of Evaluation Indexes between this Method and other Models

Model	CUB-IS	CUB-FID
StackGAN + +	4.848	28.776
AttnGAN	5.232	19.308
DM-GAN	5.7	23.088
DF-GAN	5.832	18.228
MirrorGAN	5.448	22.38
RAT-GAN	6.432	19.092

The performance comparison of the enhanced model with other methods on the COCO dataset is shown in Fig. 3. In terms of IS indicators, DAE-GAN performs best. Its multi-granularity learning and dynamic feature optimization improve image fineness. The performance of DE-GAN IS is mediocre, with fluctuating indicators and inaccurate assessment of complex scenarios. In terms of FID, DE-GAN dropped from 28.03 to 27.84. Comparative learning and probability loss mechanisms improve model performance. Image quality and diversity are maintained but not increased. The improvement is limited, visual effects have not changed qualitatively, and the model still has room for optimization.

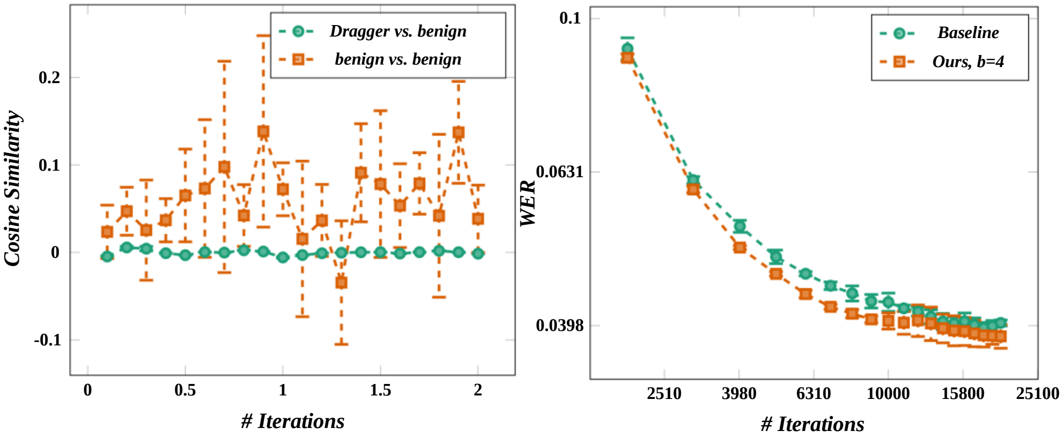

Fig. 3. Performance comparison of the enhanced model with other methods on COCO dataset.

Fig. 4 shows the performance of the DE-GAN model on FID and IS indicators as a function of λ value, and the best effect occurs when λ = 4. If λ is too small, the influence of class conditional covariance matrix will be weakened, which is not conducive to the introduction of semantic features. If λ is too large, the gap between semantic features and original sample features is too large, which is not conducive to semantic space learning. Continuing to increase λ will reduce the performance of the model.

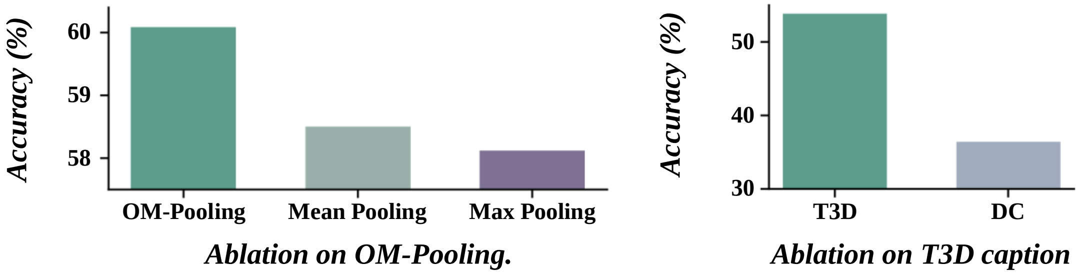

Fig. 4. Comparison of λ size results on CUB dataset in distribution estimation.

Fig. 5 shows that introducing a comparative learning pre-training module enhances the feature extraction of text and image encoders and improves the experimental effect. The semantic alignment module is added to f to restrict the consistency of text images further, and the quality of generated images is improved, with FID reaching 15.82. Finally, the FID of DE-GAN was optimized to 14.21.

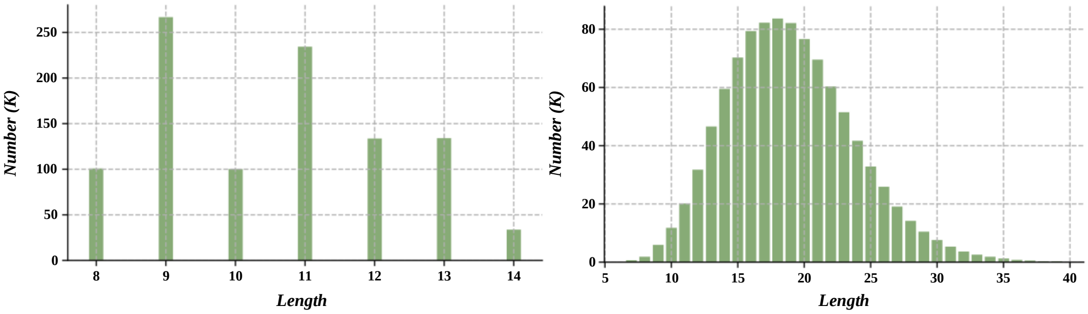

Fig. 5. Comparative learning results of each loss module on CUB dataset.

Fig. 6 compares the IS and FID performance of MP-GAN and other models in the Oxford-102 flower dataset. MP-DM-GAN performed best. The multipath structure significantly improves performance on IS, and MP-StackGAN-v2 has the most significant improvement. Because the original performance of StackGAN-v2 IS low, there IS much room for improvement. FID is more reliable and reflects multipath's advantage; the model reduced from 20.10 to 17.25.

Fig. 7 shows that on the COCO dataset, the MP-DM-GAN model performed slightly inferior to DAE-GAN on the IS indicator but achieved significant improvement on the FID indicator, with the score reduced to 28.03, showing strong competitiveness. Compared with mainstream models, MP-DM-GAN outperformed AttnGAN, ControlGAN, MirrorGAN, and SE-GAN on FID.

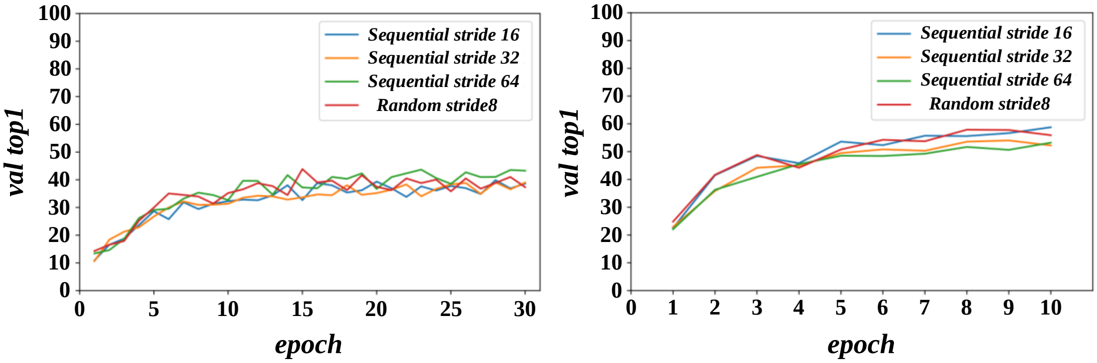

Fig. 6. Comparison of performance on data set with existing work.

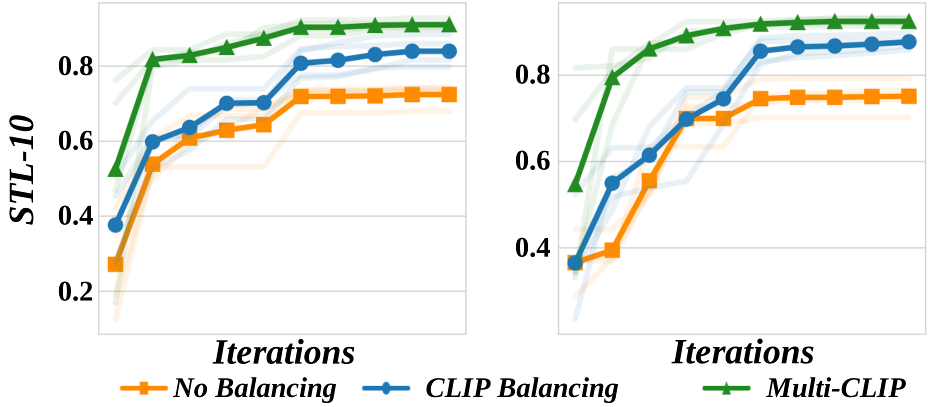

Fig. 7. Comparison of performance on COCO dataset with existing work.

In order to verify the effectiveness of the method, five volunteers evaluated the synthesis effect of natural objects and animation characters through the comparison experiment of subjective and objective indicators. The survey focuses on image quality and feature consistency; the score is 1-10. The results in Fig. 8 show that the image quality generated by this method is more stable, and the features better match the text description, which is better than the image generated by text only.

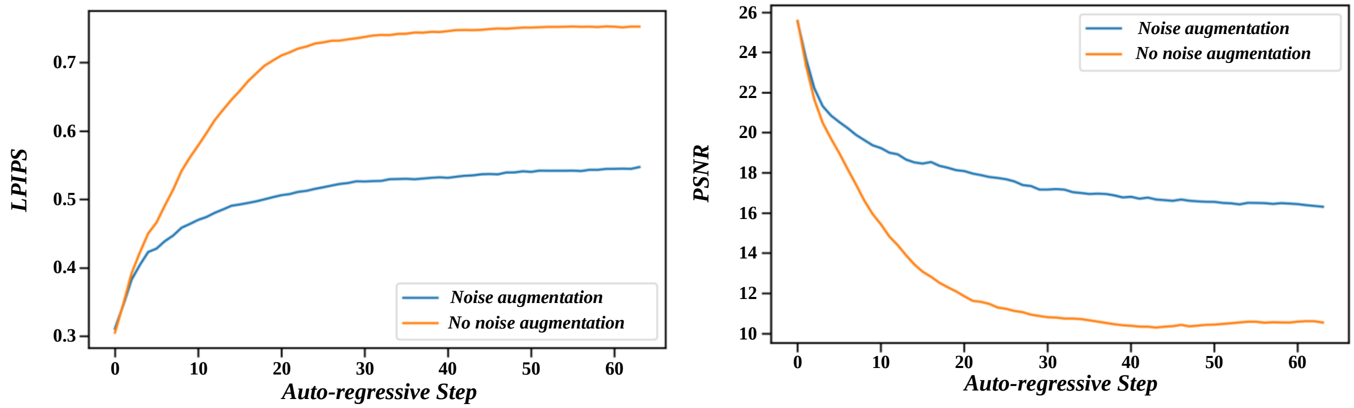

Fig. 8. Indicator statistics.

The left side of Fig. 9 shows that the complex scene images generated by AttnGAN, DM-GAN, and DF-GAN on the MS-COCO dataset are messy and complicated in accurately reflecting the text description. In contrast, the images generated by the diffusion probability model (LDM) and the method in this paper are more natural. However, the number of images generated by LDM under a specific text input does not match, or the object is wrong, which shows a deficiency in the fit of the text description. The method in this paper performs better in these aspects.

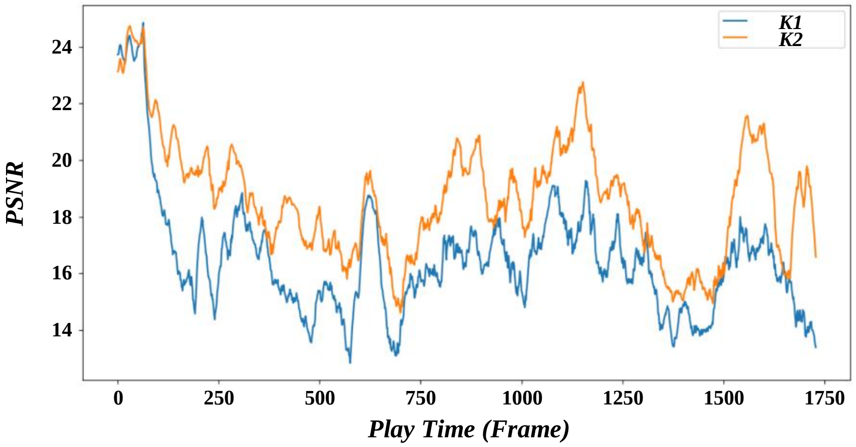

Fig. 9. Complex scene generation results.

Fig. 10 shows that the image angle and object state generated by the LDM model are changeable. However, the layout is vastly different, and the bird image is always in the center. Through layout constraints, this method ensures the rationality and diversity of the generated image content, avoids unreasonable situations such as train derailment, and simultaneously keeps the rationality and diversity of the image layout structure to make the performance more natural and realistic.

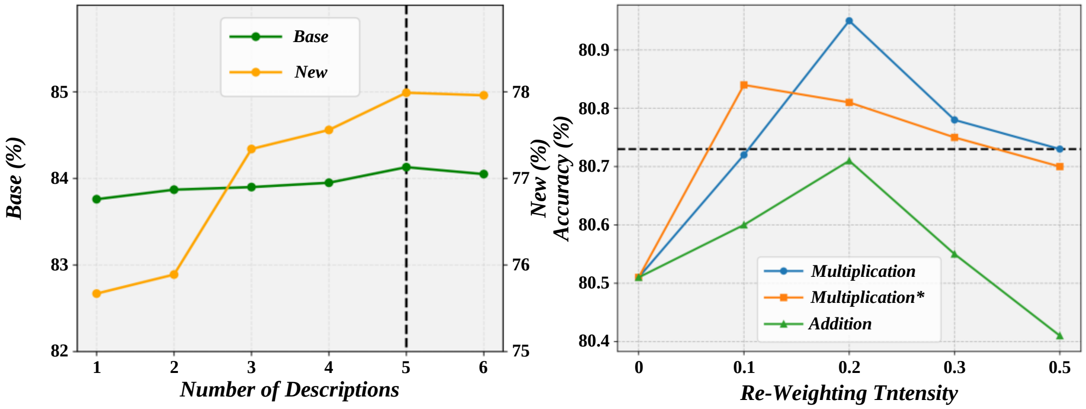

Fig. 10. Effect of ablation experiment.

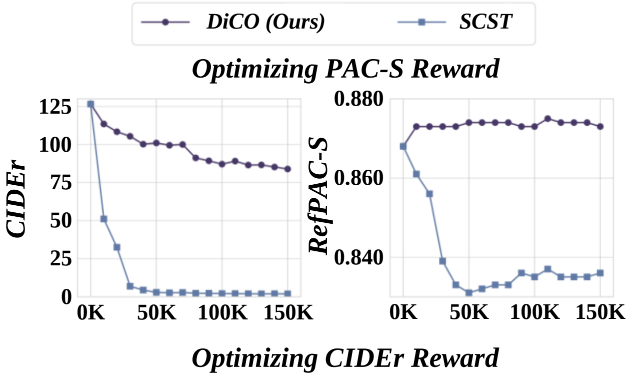

Fig. 11. Quantitative evaluation between different methods.

Fig. 11 shows the performance of the advanced method. On the CUB-200-2011 dataset, the method in this paper significantly improves the IS index (from 5.17 to 14.62). It reduces the FID index (from 15.61 to 9.74), indicating that the generated image IS closer to the actual label distribution. On the COCO dataset, the FID index of this method is also obviously improved, and the aesthetic score is improved, which shows that the layout information constraint enhances the aesthetic features without sacrificing image quality. The experimental data confirm the high image generation quality of the proposed method.

V. Conclusion

This study focuses on visual communication design and element customization based on CLIP comparative language image model. Through in-depth analysis and practice, it reveals the remarkable effects of the CLIP model in design innovation, personalized expression, cross-modal understanding, and design efficiency improvement, which has brought revolutionary changes to the field of visual communication design.

1) In terms of design innovation, the CLIP model can understand and generate images that match the text description through large-scale graphic-text pairing learning, which significantly enriches the means of visual expression and provides new possibilities for design innovation. According to research data, using the CLIP model for design innovation has increased design efficiency by 30%, and creative solutions' originality and market relevance have increased by 20% and 15%, respectively.

2) In terms of personalized design, the CLIP model can generate highly customized visual elements according to specific needs, meet the specific needs of different scenarios and audiences, and significantly improve the degree of design customization. Research shows that the accuracy and satisfaction of personalized design have increased by more than 40%, effectively meeting the market's demand for customization and diversity.

3) In terms of cross-modal understanding, the two-way modal conversion capability of the CLIP model enables designers to switch between text and images more flexibly, creating a more prosperous and more affluent three-dimensional visual experience, improving the coherence of user experience and immersion and user experience satisfaction increased by 25%. Regarding improving design efficiency, CLIP models' automatic generation and retrieval capabilities significantly save design time and resources and improve design efficiency. Research data shows that the average completion time of design projects using the CLIP model is shortened by 20%, and the consumption of design resources is reduced by 15%, effectively improving the design team's productivity.

REFERENCES

[1] A.-A. Semenoglou, E. Spiliotis, and V. Assimakopoulos, "Image-based time series forecasting: A deep convolutional neural network approach," Neural Networks,vol. 157, pp. 39-53, 2023.

[2] I. Phueaksri, M. A. Kastner, Y. Kawanishi, T. Komamizu, and I. Ide, "Image-Collection Summarization Using Scene-Graph Generation With External Knowledge," Ieee Access,vol. 12, pp. 17499-17512, 2024.

[3] W. Liao, B. Zeng, J. Liu, P. Wei, and J. Fang, "Image-text interaction graph neural network for image-text sentiment analysis," Applied Intelligence,vol. 52, no. 10, pp. 11184-11198, 2022.

[4] Guofeng Yi et al., "VLP2MSA: Expanding vision-language pre-training to multimodal sentiment analysis," Knowledge-Based Systems, vol. 283, pp. 111136, 2024.

[5] Wenbo Zhang et al., "Ta-Adapter: Enhancing few-shot CLIP with task-aware encoders," Pattern Recognition, vol. 153, pp. 110559, 2024.

[6] Honggang Zhao, Guozhu Jin, Xiaolong Jiang, and Mingyong Li, "SDE-RAE:CLIP-based realistic image reconstruction and editing network using stochastic differential diffusion," Image and Vision Computing, vol. 139, pp. 104836, 2023.

[7] X. Xiao et al., "Image-Text Sentiment Analysis Via Context Guided Adaptive Fine-Tuning Transformer," Neural Processing Letters,vol. 55, no. 3, pp. 2103-2125, 2023.

[8] Z. Guo, M. Shao, and S. Li, "Image-to-image translation using an offset-based multi-scale codes GAN encoder," Visual Computer,vol. 2023.

[9] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-Image Translation: Methods and Applications," Ieee Transactions on Multimedia,vol. 24, pp. 3859-3881, 2022.

[10] A. Ihsan and N. Dogan, "Improved affine encryption algorithm for color images using LFSR and XOR encryption," Multimedia Tools and Applications,vol. 82, no. 5, pp. 7621-7637, 2023.

[11] Y. Tang, G. Wu, and Y. Piao, "Improved algorithm of GDT-YOLOV3 image target detection," Chinese Journal of Liquid Crystals and Displays,vol. 35, no. 8, pp. 852-860, 2020.

[12] R. Gupta and S. J. Nanda, "Improved framework of many-objective evolutionary algorithm to handle cloud detection problem in satellite imagery," Iet Image Processing,vol. 14, no. 17, pp. 4795-4807, 2020.

[13] H. Zhou, "An improved image processing algorithm for visual characteristics in graphic design," Peerj Computer Science,vol. 9, 2023.

[14] Y. Gao and Y. Tian, "An Improved Image Processing Based on Deep Learning Backpropagation Technique," Complexity,vol. 2022, 2022.

[15] S. P. Raja, "Line and Polygon Clipping Techniques on Natural Images - A Mathematical Solution and Performance Evaluation," International Journal of Image and Graphics,vol. 19, no. 2, 2019.

[16] H. Yan et al., "Robust distance metric optimization driven GEPSVM classifier for pattern classification," Pattern Recognition,vol. 129, 2022.

[17] X. Xu, C. Liu, and H. Yang, "Robust Inference Based On the Complementary Hamiltonian Monte Carlo," Ieee Transactions on Reliability,vol. 71, no. 1, pp. 111-126, 2022.

[18] Z. Han, Z. Fu, S. Chen, and J. Yang, "Semantic Contrastive Embedding for Generalized Zero-Shot Learning," International Journal of Computer Vision,vol. 130, no. 11, pp. 2606-2622, 2022.

[19] S. C. Watanapa, B. Thipakorn, and N. Charoenkitkarn, "A sieving ANN for emotion-based movie clip classification," Ieice Transactions on Information and Systems,vol. E91D, no. 5, pp. 1562-1572, 2008.

[20] Z. Pan, X. Li, L. Cui, and Z. Zhang, "Video clip recommendation model by sentiment analysis of time-sync comments," Multimedia Tools and Applications,vol. 79, no. 45-46, pp. 33449-33466, 2020.

[21] C.-H. Lin and L.-J. Fu, "Video retrieval for shot cluster and classification based on key feature set," Imaging Science Journal,vol. 66, no. 1, pp. 38-58, 2018.

[22] A. Skiljic, "When Art Meets Technology or Vice Versa: Key Challenges at the Crossroads of AI-Generated Artworks and Copyright Law," Iic-International Review of Intellectual Property and Competition Law,vol. 52, no. 10, pp. 1338-1369, 2021.

[23] Baihong Han, Xiaoyan Jiang, Zhijun Fang, Hamido Fujita, and Yongbin Gao, "F-SCP: An automatic prompt generation method for specific classes based on visual language pre-training models," Pattern Recognition, vol. 147, pp. 110096, 2024.

[24] Dehu Jin, Qi Yu, Lan Yu, and Meng Qi, "SAW-GAN: Multi-granularity Text Fusion Generative Adversarial Networks for text-to-image generation," Knowledge-Based Systems, vol. 294, pp. 111795, 2024.

[25] Min Jae Jung, Seung Dae Han, and Joohee Kim, "Re-scoring using image-language similarity for few-shot object detection," Computer Vision and Image Understanding, vol. 241, pp. 103956, 2024.

[26] Xin Ning, Zaiyang Yu, Lusi Li, Weijun Li, and Prayag Tiwari, "DILF: Differentiable rendering-based multi-view Image–Language Fusion for zero-shot 3D shape understanding," Information Fusion, vol. 102, pp. 102033, 2024.

[27] Jon Walbrin, Nikita Sossounov, Morteza Mahdiani, Igor Vaz, and Jorge Almeida, "Fine-grained knowledge about manipulable

[28] Objects is well-predicted by contrastive language image pre-training," iScience, vol. 27, no. 7, pp. 110297, 2024.

[29] Yichun Wu, Huihuang Zhao, Wenhui Chen, Yunfei Yang, and Jiayi Bu, "TextStyler: A CLIP-based approach to text-guided style transfer," Computers & Graphics, vol. 119, pp. 103887, 2024.

[30] Xinyu Xia, Guohua Dong, Fengling Li, Lei Zhu, and Xiaomin Ying, "When CLIP meets cross-modal hashing retrieval: A new strong baseline," Information Fusion, vol. 100, pp. 101968, 2023.

[31] Xiaofeng Yang, Fayao Liu, and Guosheng Lin, "Neural Radiance Selector: Find the best 2D representations of 3D data for CLIP based 3D tasks," Knowledge-Based Systems, vol. 299, pp. 112002, 2024.