DECIDER: Difference-aware Contrastive Diffusion Model with Adversarial Perturbations for Image Change Captioning

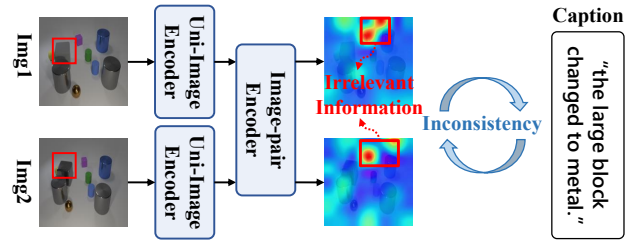
Guojin Zhong[1*], Jinhong Hu[1, 2*], Jiajun Chen[3], Jin Yuan[1†], Wenbo Pan[4†]

[1]College of Computer Science and Electronic Engineering, Hunan University
[2]GuangDong Engineering Technology Research Center of Intelligent Service of Urban and Rural Planning and Construction
[3]School of Robotics, Hunan University
[4]CRRC Zhuzhou Institute
gjzhong@hnu.edu.cn, jinhonghu@hnu.edu.cn, chenjiajun@hnu.edu.cn, yuanjin@hnu.edu.cn, panwb@mail.ustc.edu.cn

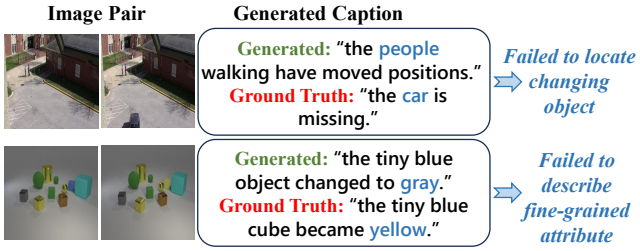
Abstract

Image change captioning (ICC) poses great challenges stemming from describing subtle differences between two similar images in natural language, significantly increasing the complexity of feature extraction and cross-modal learning compared to the image captioning task. Existing ICC methods often suffer from two key challenges: 1) Massive irrelevant information of uni-image features leads to suboptimal visual difference representations; 2) Imprecise inter-modality correspondence degrades the quality of generated captions. This paper proposes a **D**ifference-awar**e** **C**ontrastive **D**iffusion Model with A**d**versarial **Per**turbations (DECIDER) for ICC due to the excellent performance of diffusion models in image/text generation. Technically, difference-aware cross-modal learning is developed to suppress irrelevant information and learn compact yet robust visual difference representations. This is achieved by optimizing a novel objective mathematically derived from the information bottleneck principle that excels in filtering redundant features and highlighting differences. Furthermore, we propose to dynamically generate "hard" positive and negative samples via adversarial perturbations, which are involved in contrastive diffusion training with a tighter variational bound. This design encourages our DECIDER to excavate and construct complex correspondences between visual differences and captions, thereby improving generalization performance. Extensive experiments on four datasets demonstrate that DECIDER significantly exceeds state-of-the-art performance.

Code — https://github.com/zgj77/DECIDER



(a) Irrelevant information in visual difference representation



(b) Incorrect inter-modality correspondence

Figure 1: Examples to show two ICC challenges: one is the presence of irrelevant information in visual difference representations, and the other is the failure of correspondence in matching generated captions with input images.

Introduction

The natural modality gap between vision and text poses great challenges for machines to understand visual information by using textual descriptions and promotes the birth of Image Captioning (IC) (Vinyals et al. 2015), which aims at generating accurate linguistic captions for an image. Beyond IC, Image Change Captioning (ICC) (Jhamtani and Berg-Kirkpatrick 2018) involves an in-depth comparison of two similar images and describes their tiny differences in natural language, raising significant technical challenges. Even large vision-language models like Qwen-VL (Bai et al. 2023) and LLaVA (Liu et al. 2024), which excel in various multi-modal tasks, struggle with ICC task attributed to their limited ability to accurately model visual differences. Since ICC has high practical values in real-world applications like reporting industrial anomalies, detecting and describing lesions, it is crucial to develop an effective model for ICC.

Most existing ICC methods employ the encoder-decoder framework to uniformly model the complex cross-modal correlations between visual differences and linguistic representations. The visual encoder specifically aims to capture the subtle differences between two images (Park, Darrell, and Rohrbach 2019), while the linguistic decoder produces descriptions accordingly (Kim et al. 2021). With the advent of transformers (Vaswani et al. 2017), recent studies

*These authors contributed equally.
†Corresponding author.

(Yao, Wang, and Jin 2022) introduce transformers with several auxiliary losses into the ICC task. Despite their success, these autoregressive models usually suffer from the error accumulation problem (Gu et al. 2022). Alternatively, non-autoregressive diffusion models have recently shown superior performance in text/image generation (Li et al. 2022b) due to their advantages in preventing mode collapse (Zhong et al. 2023a) and training instability. In the realm of image captioning, it has been demonstrated that diffusion models, which erode discrete text tokens and then restore them (Zhu et al. 2022), can generate accurate captions by effectively exploring the complex inter-modality correlations.

Motivated by this, we attempt to design a diffusion model for ICC by exploring its key challenges. First, unlike the comprehensive description of an image required in the IC task, ICC focuses on describing local regions with subtle changes between two similar images. This increases the difficulty of encoding visual differences, as the vast amount of irrelevant information in uni-image features would inevitably interfere with the encoding process (see Fig. 1 (a)). Second, the generated captions sometimes fail to locate changing objects or to describe fine-grained attributes, *etc.*, which illustrates that the complex correspondence between captions and visual differences is lost due to the inter-modality heterogeneity (see Fig. 1 (b)).

To tackle the above issues, we propose a **D**ifference-awar**e** **C**ontrastive D**i**ffusion with A**d**versarial P**er**turbations (DECIDER) for ICC. DECIDER incorporates difference-aware cross-modal learning based on the information bottleneck principle to model compact yet robust visual difference representations and dynamically produces "hard" contrastive samples for diffusion training via adversarial perturbations to enhance the correspondence between visual difference and change description. Concretely, given a pair of images with the corresponding description, we first employ the vision and language encoders of CLIP to encode them into primary visual difference and textual representations, respectively. On this basis, we perform the proposed difference-aware cross-modal learning whose objective is mathematically derived from the information bottleneck principle to suppress irrelevant information of primary visual difference representation, *i.e.*, minimizing the mutual information between it and the input images, while maximizing the mutual information between it and the change captions, thereby facilitating precise localization of regions with subtle differences. Afterward, we dynamically generate the "hard" contrastive samples via adversarial perturbations, where the positive samples have distant representations but similar semantics, and the negative samples have close representations but different semantics. Compared to naive contrastive samples, these "hard" contrastive samples provide a more robust supervised signal for cross-modal learning. Finally, they are both incorporated into the diffusion model to establish a tighter evidence boundary and enhance the learning of correspondences between visual differences and change captions. Extensive experiments demonstrate that our DECIDER significantly outperforms the state-of-the-art (SOTA) methods. The main contributions are summarized as follows:

- Our study deeply analyzes the key challenges of the ICC task and designs a novel difference-aware contrastive diffusion model with adversarial perturbations (DECIDER) for generating precise change captions. DECIDER is underpinned by rigorous mathematical theory.
- We propose difference-aware cross-modal learning mathematically derived from the information bottleneck principle to facilitate inter-modality alignment, yielding compact yet robust visual difference representations that enhance the accuracy of caption generation.
- We propose a contrastive diffusion model that incorporates "hard" contrastive samples dynamically generated via adversarial perturbations, achieving a tighter optimization bound. To the best of our knowledge, it is the first work to introduce adversarial strategies into text diffusion training to improve generalization performance.

Related Work

Image Change Captioning

Image change captioning (ICC) aims to articulate the differences present in a pair of similar images by using natural language (Jhamtani and Berg-Kirkpatrick 2018). Previous research has mainly focused on the improvement of fine-grained feature extraction in a single modality like CUDA (Park, Darrell, and Rohrbach 2019), VAM (Shi et al. 2020), and VACC (Kim et al. 2021). For instance, (Huang et al. 2021) and (Tu et al. 2021) design convolution-based visual feature encoders to capture viewpoint information between image pairs and then feed them into RNN (Li et al. 2018, 2022a) to generate change captions. Inspired by the development of transformers (Sun 2024; Cheng and Sun 2024), (Yao, Wang, and Jin 2022) and (Tu et al. 2023a) propose different pre-training losses for transformers to learn the mapping between text and visual changes. To tackle the impact of pseudo-differences (Sun et al. 2024) under viewpoint changes (Yue et al. 2024), (Tu et al. 2023b) exploits self-supervised learning to extract invariant representations, (Tu et al. 2024) designs a syntax-calibrated multi-aspect relation transformer to learn effective change features. Despite the success, these autoregressive approaches above usually suffer from severe error accumulation, which affects the quality of the generated text. Still, these approaches do not effectively alleviate the effect of redundant visual information, leading to insufficient learning of visual difference representations. Instead, we propose a contrastive diffusion model with difference-aware cross-modal learning to eliminate irrelevant visual features and generate accurate change captions in a non-autoregressive manner.

Diffusion Model for Text Generation

Since diffusion models (DMs) (Ho, Jain, and Abbeel 2020; Zhong et al. 2024) have achieved significant success in audio and image generation (Ruan et al. 2023), researchers recently attempted to overcome the constraints imposed by the discreteness of text and apply DM to text generation tasks such as sequence-to-sequence and image captioning. The existing works on text generation mainly follow two directions. Firstly, (Hoogeboom et al. 2021) gradually adds categorical noise instead of Gaussian noise to learn

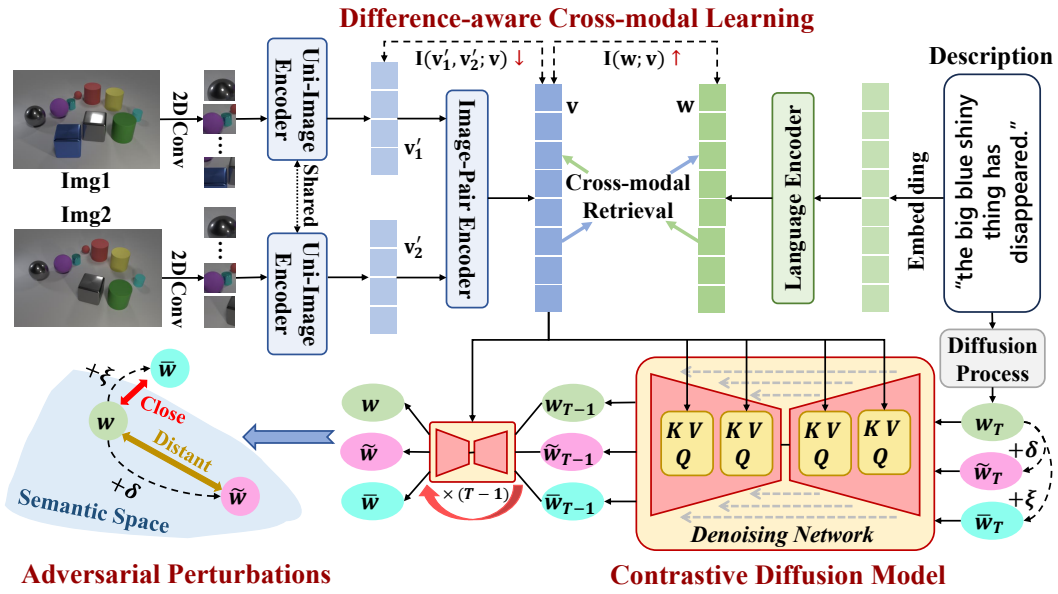


Figure 2: The DECIDER framework incorporates the difference-aware cross-modal learning that enhances both the language and visual difference encoders, along with a diffusion model augmented by adversarial perturbations. Within this framework, $\mathbf{v}'_*$ and $\mathbf{v}$ represent the uni-image primary representation and the visual difference representation, respectively. $\mathbf{w}$, $\tilde{\mathbf{w}}$ and $\overline{\mathbf{w}}$ denote the textual representations for the anchor, positive and negative samples, respectively.

the character-level text generation in the denoising process. (Savinov et al. 2021) and (Austin et al. 2021) design absorbing states ([MASK]) and transition matrices to corrupt and refine the text token. However, these kinds of coarse-grained corruptions inevitably affect the semantic expression and coherence of sentence (Gao et al. 2022). In contrast, (Chen, Zhang, and Hinton 2022) and (Luo et al. 2023) encode the discrete text into binary bits. (Li et al. 2022b), (Gong et al. 2023b), and (Lovelace et al. 2024) deploy continuous DMs on a discrete text by utilizing embedding and rounding. Motivated by these works, we introduce the DM for ICC and leverage positive and negative samples to improve the generalization performance.

Contrastive Diffusion Model

To enhance the connection between targets and conditional variables, (Zhu et al. 2023) proposes a conditional discrete contrastive diffusion loss, which incorporates negative samples into the training process of DM. On this basis, (Zhong et al. 2023b) explicitly introduces both positive and negative samples into DM to provide a tighter bound for evidence and derive a novel variational objective. Different from them, our proposed contrastive diffusion model exploits adversarial perturbations to dynamically produce "hard" contrastive samples, providing a more robust supervised signal for cross-modal learning and generating accurate captions.

Methodology

Fig. 2 presents our **D**ifference-awar**e** **C**ontrastive **D**iffusion Model with A**d**versarial **Per**turbations (DECIDER) for ICC, which consists of three parts: 1) The language and vision encoders convert input caption and image-pair into textual and visual difference representations, respectively. 2) The proposed difference-aware cross-modal learning promotes the excavation of image difference features and the alignment of bi-modal representations. 3) The proposed contrastive diffusion model with adversarial perturbations generates sentences describing the differences between two images.

Input Representation

The input representation module involves a vision encoder to model the visual differences between two images (I_1, I_2), and a language encoder to represent their change caption S.

Text Encoding Given a caption S with the length M, we first tokenize each word w_m and then encode them into textual representations $\mathbf{w} \in \mathbb{R}^{M\times D}$ with dimension D by employing the language encoder $\mathcal{E}_L$ consisting of an embedding transformation $f_\phi(\cdot)$ and N_l transformer layers as:

$$\mathbf{w} = \mathcal{E}_L(S) = \mathcal{E}_L(\{f_\phi(w_1), f_\phi(w_2), ..., f_\phi(w_M)\}). \quad (1)$$

Visual Difference Modeling Given a pair of input images (I_1, I_2), the CLIP model with excellent zero-shot performance is introduced to convert each image into K patches and extract primary visual representations $\mathbf{v}'_1, \mathbf{v}'_2 \in \mathbb{R}^{K\times D}$ via a siamese uni-image encoder $\mathcal{E}_s$ consisting of N_s transformer layers. Then they are fed into a image-pair encoder $\mathcal{E}_p$ containing N_p transformer layers to obtain primary visual difference representation $\mathbf{v} \in \mathbb{R}^{2K\times D}$, denoted as:

$$\mathbf{v} = \mathcal{E}_p([\mathcal{E}_s(I_1); \mathcal{E}_s(I_2)]), \quad (2)$$

where $[\cdot;\cdot]$ denotes concatenation of two vectors.

Difference-aware Cross-modal Learning via Information Bottleneck

Since extracted visual features contain redundant information that is irrelevant to the change caption, previous methods often fail to accurately encode visual changes between two images. Inspired by the superiority of the Information Bottleneck (IB) principle (Jiang, Liu, and Zheng 2023) in encoding compact yet informative features, we propose a novel approach termed difference-aware cross-modal learning to adapt pre-trained visual difference representations for captioning, effectively locating the changing regions.

For vision encoding, we seek a robust visual difference representation $\mathbf{v}$ that can be maximally predictive of the target caption $\mathbf{w}$ for the ICC task, while eliminating redundant information from uni-image representations $\mathbf{v}'_1$ and $\mathbf{v}'_2$ to prevent the model from focusing on irrelevant regions. Formally, this difference-aware learning (DAL) is formulated by IB as an information trade-off, which aims to maximize the Lagrangian objective to find the optimal representation:

$$\mathcal{L}_{dal} = I(\mathbf{w};\mathbf{v}) - \sigma I(\mathbf{v}'_1, \mathbf{v}'_2; \mathbf{v}), \tag{3}$$

where $\sigma \geq 0$ is a scalar that controls the weight of information compression, $I(\cdot;\cdot)$ denotes mutual information. The first term in Eq. (3) enhances the correlations between target $\mathbf{w}$ and visual difference representation $\mathbf{v}$, while the second term facilitates $\mathbf{v}$ in obtaining change-relevant information from uni-image representation $\mathbf{v}'_1$ and $\mathbf{v}'_2$.

For $I(\mathbf{w};\mathbf{v})$, we can rewrite it by leveraging the conditional probability with BA lower bound (2004):

$$\begin{aligned} I(\mathbf{w};\mathbf{v}) &= \int p(\mathbf{w},\mathbf{v}) \log \frac{p(\mathbf{w}|\mathbf{v})}{p(\mathbf{w})} d\mathbf{w} d\mathbf{v} \\ &\geq \int p(\mathbf{w},\mathbf{v}) \log q(\mathbf{w}|\mathbf{v}) d\mathbf{w} d\mathbf{v} - \int p(\mathbf{w}) \log p(\mathbf{w}) d\mathbf{w} \\ &= -E(\mathbf{w}|\mathbf{v}) + E(\mathbf{w}), \end{aligned} \tag{4}$$

where $E(\cdot)$ denotes entropy and $q(\mathbf{w}|\mathbf{v})$ is an auxiliary distribution for $p(\mathbf{w}|\mathbf{v})$. The target entropy term $E(\mathbf{w})$ can be ignored because it is independent of the optimization process, in which case maximizing the lower bound of $I(\mathbf{w};\mathbf{v})$ is equivalent to minimizing the cross-entropy $E(\mathbf{w}|\mathbf{v})$. And we can utilize a symmetric retrieval loss between image-pair and caption to equivalent bound maximization of $I(\mathbf{w};\mathbf{v})$:

$$\begin{aligned} \hat{I}(\mathbf{w};\mathbf{v}) = &- \frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\mathbf{v}^{(i)\top} \cdot \mathbf{w}^{(i)})}{\sum_{j=1}^{B} \exp(\mathbf{v}^{(i)\top} \cdot \mathbf{w}^{(j)})} \\ &- \frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\mathbf{w}^{(i)\top} \cdot \mathbf{v}^{(i)})}{\sum_{j=1}^{B} \exp(\mathbf{w}^{(i)\top} \cdot \mathbf{v}^{(j)})}, \end{aligned} \tag{5}$$

where B denotes the number of samples within a training batch. Eq. (5) encourages our model to learn invariant and discriminative features by treating paired $\mathbf{v}^{(i)}$ and $\mathbf{w}^{(i)}$ from the same sample (i) as similar while pushing apart unpaired samples. For $I(\mathbf{v}'_1, \mathbf{v}'_2; \mathbf{v})$ in Eq. (3), we consider its upper bound and then derive a tighter lower bound for DAL following Theorem 1 (see Supplementary Material for details).

Theorem 1. ***(Lower Bound of DAL.)*** *Given two uni-image representations $\mathbf{v}'_1$ and $\mathbf{v}'_2$, the visual difference representation $\mathbf{v}$ obtained by deterministic process $\mathbf{v} = \mathcal{E}_p([\mathbf{v}'_1; \mathbf{v}'_2])$. The difference-aware learning can be bounded as:*

$$\mathcal{L}_{dal} \geq I(\mathbf{w};\mathbf{v}) - \sigma[I(\mathbf{v}'_1;\mathbf{v}) + I(\mathbf{v}'_2;\mathbf{v}) + D_{SKL}], \tag{6}$$

where D_{SKL} denotes the symmetric Kullback-Leibler divergence that is the average of divergences $D_{KL}(p(\mathbf{v}|\mathbf{v}'_1)||p(\mathbf{v}|\mathbf{v}'_2))$ and $D_{KL}(p(\mathbf{v}|\mathbf{v}'_2)||p(\mathbf{v}|\mathbf{v}'_1))$.

Notably, for the mutual information $I(\mathbf{v}'_*;\mathbf{v})$ in Theorem 1 where $*$ indicates uni-image index 1 or 2, we estimate them in a sample-based differentiable manner (Cheng et al. 2020) since the conditional distributions $p(\mathbf{v}^{(i)}|\mathbf{v}'^{(i)}_*)$ are tractable:

$$\hat{I}(\mathbf{v}'_*;\mathbf{v}) = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}[\log p(\mathbf{v}^{(i)}|\mathbf{v}'^{(i)}_*) - \log p(\mathbf{v}^{(j)}|\mathbf{v}'^{(i)}_*)]. \tag{7}$$

Eq. (7) facilitates the visual difference $\mathbf{v}$ more compact and robust by optimizing the conditional distributions of $\mathbf{v}$ given $\mathbf{v}'_*$. Overall, benefiting from the proposed difference-aware cross-modal learning, we significantly suppress the redundant information in visual difference representations and facilitate our input encoder to learn the inter-modality alignment between vision and language for the ICC task.

Contrastive Diffusion Model with Adversarial Perturbations

Most DMs first map a discrete caption into a continuous text embedding $\mathbf{w}$ and then corrode it by adding Gaussian noise via Markov chains. Then DMs learn the correspondence between $\mathbf{w}$ and visual difference $\mathbf{v}$ by optimizing the variational lower bound loss via the denoising process:

$$\begin{aligned} \mathcal{L}_{vlb}(\mathbf{w},\mathbf{v}) = &\mathbb{E}_{q(\mathbf{w}_{1:T}|\mathbf{w}_0)}[\log \frac{q(\mathbf{w}_T|\mathbf{w}_0)}{p_\theta(\mathbf{w}_T)} - \log p_\theta(\mathbf{w}|\mathbf{w}_0,\mathbf{v}) \\ &+ \sum_{t=2}^{T} \log \frac{q(\mathbf{w}_{t-1}|\mathbf{w}_0,\mathbf{w}_t)}{p_\theta(\mathbf{w}_{t-1}|\mathbf{w}_t,\mathbf{v})} + \log \frac{q(\mathbf{w}_0|\mathbf{w})}{p_\theta(\mathbf{w}_0|\mathbf{w}_1,\mathbf{v})}], \end{aligned} \tag{8}$$

where $q(\cdot)$ and $p_\theta(\cdot)$ denote the forward distribution and approximation of posterior distribution, respectively.

On this basis, we seek to further enhance the correlation between $\mathbf{v}$ and generated text $\mathbf{w}$ by introducing contrastive learning into our diffusion model inspired by (Zhu et al. 2023; Zhong et al. 2023b). Specifically, the training process involves both positive sample $\widetilde{\mathbf{w}}$ and negative samples $\overline{\mathbf{w}}$, which considers a novel optimization objective, *i.e.*, maximizing the logarithmic joint conditional likelihood $\log p(\mathbf{w}, \widetilde{\mathbf{w}}|\mathbf{v})$ while minimizing $\log p(\overline{\mathbf{w}}|\mathbf{v})$:

$$maximize \ \log p(\mathbf{w}, \widetilde{\mathbf{w}}|\mathbf{v}) - \lambda \log p(\overline{\mathbf{w}}|\mathbf{v}), \tag{9}$$

where λ is a balanced weight. However, it is difficult to directly optimize this objective, thus we instead seek a tractable lower bound for it by calculating the evidence lower bound (ELBO) of $\log p(\mathbf{w}, \widetilde{\mathbf{w}}|\mathbf{v})$ and evidence upper bound (EUBO) of $p(\overline{\mathbf{w}}|\mathbf{v})$:

$$\begin{aligned} &\log p(\mathbf{w}, \widetilde{\mathbf{w}}|\mathbf{v}) - \lambda \log p(\overline{\mathbf{w}}|\mathbf{v}) \\ \geq &\underbrace{\text{ELBO}(\log p(\mathbf{w}, \widetilde{\mathbf{w}}|\mathbf{v})) - \lambda \, \text{EUBO}(\log p(\overline{\mathbf{w}}|\mathbf{v}))}_{\text{a tractable lower bound of objective}}. \end{aligned} \tag{10}$$

The effectiveness of ELBO and EUBO lies in their ability to find a tractable lower bound for optimizing the objective. By exposing the diffusion model to contrastive samples, the correspondences between captions and visual differences are effectively explored. We provide proof that their mutual information lower bound is increased from the perspective of information theory, which can improve the accuracy of correspondences (see Supplementary Material for details).

However, selecting "hard" contrastive samples is nontrivial. First, there is a lack of general yet effective augmentation strategies in the ICC task to produce positive samples that are semantically consistent with the target text. Moreover, negative samples randomly sampled within the same batch may be far away from the target text in representation space, which leads to suboptimal training. To tackle these issues, we add adversarial perturbations (Lee, Lee, and Hwang 2020) to $\mathbf{w}$, dynamically generating "hard" contrastive samples $\widetilde{\mathbf{w}}$ and $\overline{\mathbf{w}}$ to achieve the objective in Eq. (10).

ELBO with Hard Positive Sample Concretely, the logarithmic joint conditional likelihood $\log p(\mathbf{w}, \widetilde{\mathbf{w}}|\mathbf{v})$ is model by considering diffusion variables $\mathbf{w}_t$/$\widetilde{\mathbf{w}}_t$ at each timestep t (see Supplementary Material for detailed proof):

$$\begin{aligned}\log p(\mathbf{w}, \widetilde{\mathbf{w}}|\mathbf{v}) =& \log \int_{\widetilde{\mathbf{w}}_t} \int_{\mathbf{w}_t} p(\mathbf{w}_0, \mathbf{w}_t, \widetilde{\mathbf{w}}_0, \widetilde{\mathbf{w}}_t|\mathbf{v}) d\mathbf{w}_t d\widetilde{\mathbf{w}}_t \\ \geqslant& \mathcal{L}_{vlb}(\mathbf{w}, \mathbf{v}) + \mathcal{L}_{vlb}(\widetilde{\mathbf{w}}, \mathbf{v}) - \mathbb{E}_t[I(\mathbf{w}; \widetilde{\mathbf{w}})].\end{aligned} \tag{11}$$

The "hard" positive sample $\widetilde{\mathbf{w}}$ is produced by adding a perturbation $\boldsymbol{\delta}$ to $\mathbf{w}$, which is expected to be semantically consistent with $\mathbf{w}$ while being far away from $\mathbf{w}$ in the embedding space. Unfortunately, it is intractable to directly compute $\boldsymbol{\delta}$, thus we approximate it in two steps. First, we find the gradient direction $-\mathbf{g}_1$ away from $\mathbf{w}$ by calculating InfoNCE-form contrastive loss (He et al. 2020), and then obtain an intermediate variable $\ddot{\mathbf{w}}$ distant from $\mathbf{w}$ using gradient sign approach (Goodfellow, Shlens, and Szegedy 2014):

$$\mathcal{L}_{info} = \log \frac{\exp(\mathbf{w}^\top \cdot \widetilde{\mathbf{w}}_{na}/\tau)}{\sum_{\overline{\mathbf{w}}_{na} \in G} \exp(\mathbf{w}^\top \cdot \overline{\mathbf{w}}_{na}/\tau)}, \tag{12}$$

$$\ddot{\mathbf{w}} = \mathbf{w} - \eta \frac{\mathbf{g}_1}{||\mathbf{g}_1||_2}, \text{ where } \mathbf{g}_1 = \nabla_{\mathbf{w}} \mathcal{L}_{info}, \tag{13}$$

where τ is a temperature coefficient. $\widetilde{\mathbf{w}}_{na}$ and $\overline{\mathbf{w}}_{na}$ are naive contrastive samples, where the former is calculated by passing $\mathbf{w}$ through the Gaussian noise-augmented embedding layer (Gong et al. 2023a), the latter is randomly selected. Afterward, the KL-divergence between $p(\mathbf{w}|\mathbf{v})$ and $p(\ddot{\mathbf{w}}|\mathbf{v})$ is minimized to maintain the semantics of $\widetilde{\mathbf{w}}$:

$$\mathcal{L}_{KL} = \mathrm{D}_{KL}(p(\mathbf{w}|\mathbf{v})||p(\ddot{\mathbf{w}}|\mathbf{v})), \tag{14}$$

$$\widetilde{\mathbf{w}} = \ddot{\mathbf{w}} - \eta \frac{\mathbf{g}_2}{||\mathbf{g}_2||_2}, \text{ where } \mathbf{g}_2 = \nabla_{\ddot{\mathbf{w}}} \mathcal{L}_{KL}. \tag{15}$$

Since $\mathbf{g}_1$ and $\mathbf{g}_2$ point in the directions of the steepest ascent for Eq. (12) and (14), respectively, adding adversarial perturbations in the opposite directions can generate "hard" positive samples that are semantically consistent but have distant representations. This allows our model to more effectively capture the essential characteristics of similar samples.

EUBO with Hard Negative Sample The logarithmic conditional likelihood $\log p(\overline{\mathbf{w}}|\mathbf{v})$ is minimized based on EUBO theorem (see Supplementary Material) to provide a tighter bound for variational evidence (Dieng et al. 2017):

$$\log p(\overline{\mathbf{w}}|\mathbf{v}) \leq \frac{1}{2} \log \mathbb{E}_{q(\overline{\mathbf{w}}_t)}[(\frac{p(\overline{\mathbf{w}}_0, \overline{\mathbf{w}}_t)}{q(\overline{\mathbf{w}}_t|\mathbf{v})})^2] \triangleq e^{\mathcal{L}_{vlb}(\overline{\mathbf{w}}_0, \mathbf{v})}. \tag{16}$$

Contrary to $\widetilde{\mathbf{w}}$, the "hard" negative sample $\overline{\mathbf{w}}$ produced by adding a tiny perturbation $\boldsymbol{\xi}$ whose norm is limited within γ has representations close to $\mathbf{w}$. At the same time, its semantics are expected to be different from $\mathbf{w}$, *i.e.*, its conditional likelihood about $\mathbf{v}$ remains low:

$$\overline{\mathbf{w}} = \mathbf{w} + \boldsymbol{\xi}, \text{ s.t. } \boldsymbol{\xi} = \underset{\boldsymbol{\xi}, ||\boldsymbol{\xi}||_2 \leq \gamma}{\arg\min} \log p(\mathbf{w} + \boldsymbol{\xi}|\mathbf{v}). \tag{17}$$

Notably, solving $\boldsymbol{\xi}$ exactly is intractable likewise. We find the gradient direction $-\mathbf{g}_3$ that changes the semantics of $\mathbf{w}$ by calculating conditional likelihood $\log p(\mathbf{w}|\mathbf{v})$ and approximately implement Eq. (17), denoted as:

$$\overline{\mathbf{w}} = \mathbf{w} - \gamma \frac{\mathbf{g}_3}{||\mathbf{g}_3||_2}, \text{ where } \mathbf{g}_3 = \nabla_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{v}). \tag{18}$$

This approach facilitates our model to learn the discriminative information within different samples, aligning change descriptions and visual differences.

Training Scheme

The optimization of our DECIDER includes two parts: Difference-aware Cross-modal Learning and Contrastive Diffusion Model training. We first maximize the objective $\mathcal{L}_{dal}$ in Eq. (6) to adapt the visual difference representation for change captioning. Next, the final objectives of contrastive diffusion model can be expressed as:

$$minimize\, \mathcal{L}_{vlb}(\mathbf{w}, \mathbf{v}) + \mathcal{L}_{vlb}(\widetilde{\mathbf{w}}, \mathbf{v}) - \mathbb{E}_t[I(\mathbf{w}; \widetilde{\mathbf{w}})] - \lambda e^{\mathcal{L}_{vlb}(\overline{\mathbf{w}}, \mathbf{v})}, \tag{19}$$

here we adopt the contrastive loss with adversarial samples at timestep t to approximate the mutual information term (see Eq. (12), where $\widetilde{\mathbf{w}}$ and $\overline{\mathbf{w}}$ replace $\widetilde{\mathbf{w}}_{na}$ and $\overline{\mathbf{w}}_{na}$, respectively). The whole training algorithms can be found in Algorithm 1 and 2 of the Supplementary Material.

Experiments

Experimental Setup

We conduct experiments on the following datasets: CLEVR-Change (Park, Darrell, and Rohrbach 2019), Image-Editing-Request (Tan et al. 2019), Spot-the-Change (Jhamtani and Berg-Kirkpatrick 2018), and Birds-to-Words (Forbes et al. 2019). To evaluate the performance, we employ four metrics including: BLEU-4 (B) (Papineni et al. 2002), METEOR (M) (Banerjee and Lavie 2005), CIDEr-D (C) (Vedantam, Lawrence Zitnick, and Parikh 2015), and ROUGE-L (R) (Lin 2004). The detailed implementation settings are provided in the Supplementary Material.

Comparison with State-of-the-art Methods

Results on CLEVR-Change In addition to several ICC methods, we also introduce advanced large vision language

Model	Input	CLEVR-Change				Image-Editing-Request				Spot-the-Change				Birds-to-Words			
		B	M	C	R	B	M	C	R	B	M	C	R	B	M	C	R
VARDTrans (TIP23)	ResNet	55.4	40.1	126.4	73.8	9.9	14.8	35.7	39.0	-	12.5	30.3	29.3	25.2	19.4	14.8	45.3
SCORER (ICCV23)	ResNet	56.3	41.2	126.8	74.5	10.0	15.0	33.4	39.6	10.2	12.2	38.9	-	27.6	24.6	19.2	48.7
DIRL (ECCV24)	ResNet	54.6	38.1	123.6	71.9	10.9	15.0	34.1	41.0	10.3	13.8	40.9	32.8	-	-	-	-
SMART (TPAMI24)	ResNet	56.1	**40.8**	127.0	74.2	10.4	15.1	34.6	40.3	-	13.5	39.4	31.6	24.7	22.3	17.6	46.3
Qwen-VL	Qwen-7B	48.9	36.0	119.8	71.2	8.3	12.4	30.8	36.6	9.3	10.3	34.6	29.5	26.4	22.2	20.4	45.6
LLaVA-1.5 (NeurIPS23)	Vicuna-7B	49.7	35.4	122.4	70.8	8.7	12.9	31.3	36.9	9.5	11.0	34.2	28.8	27.3	22.6	21.4	46.3
DECIDER (Ours)	ResNet	56.4	39.7	131.3	75.3	10.6	15.8	38.0	40.8	10.7	14.2	39.9	41.6	29.6	26.4	25.7	49.8
DECIDER (Ours)	CLIP	**57.2**	40.4	**134.6**	**76.5**	**11.3**	**16.7**	**39.5**	**42.4**	**11.5**	**15.4**	**40.3**	**43.2**	**30.8**	**27.6**	**28.4**	**51.5**

Table 1: Performance comparison results across four datasets.

Model	C	T	M	A	D
NCT (TMM23)	140.2	128.8	86.0	128.4	129.0
VARDTrans (TIP23)	140.6	128.5	82.5	125.2	127.4
SCORER (ICCV23)	146.2	133.7	92.2	131.1	133.9
SMART (TPAMI24)	146.0	**135.6**	90.1	129.2	136.8
Ours	**157.4**	131.2	**106.2**	**133.6**	**137.0**

Table 2: Performance comparison on different change types of CLEVR-Change including C (Color), T (Texture), M (Move), A (Add), and D (Drop) by CIDEr.

models Qwen-VL (Bai et al. 2023) and LLaVA-1.5 (Liu et al. 2024) for comprehensive comparison and use *"Describe their differences in one sentence"* as their prompts. Table 1 demonstrates the performance comparison results on the CLEVR-Change dataset. Benefiting from the diffusion model with difference-aware cross-modal learning and adversarial contrastive samples, DECIDER achieves promising performance with a significant improvement, boosting the CIDEr from 128.9 to 134.6 and the ROUGE-L from 74.5 to 76.5. Although the parameter scales of Qwen-VL and LLaVa-1.5 are much larger than that of ICC methods, their performance is significantly worse due to their weak capability in capturing tiny visual differences. Table 2 further demonstrates the results of different change types. DECIDER achieves the top CIDEr scores on C, M, A, and D with salient changes between images, and the second top scores on T with subtle changes. We analyze our approach has a certain superiority on samples with salient changes since the diffusion imposes global corrosion on a sentence instead of local word masking by the previous approaches, which is more beneficial for building a stronger correlation between change caption and salient differences.

Results on Birds-to-Words Table 1 demonstrates the evaluation results on the Birds-to-Words dataset. Due to the complex natural background and subtle differences in bird appearances, the performance deteriorates as compared to that on the CLEVR-Change. Our DECIDER achieves a significant improvement even compared to IDC-PCL with extra data, which indicates that our model can also be extended to natural scenes, effectively modeling the complex correspondence between visual differences and captions.

Results on Image-Editing-Request Deploying to the highly diverse Image-Editing-Request dataset, all the approaches also show a decrease in performance, as illustrated in Table 1. Nevertheless, DECIDER significantly outperforms previous SOTA approaches, which indicates that DECIDER can also be applied to datasets with high diversity.

Results on Spot-the-Change Extended to the Spot-the-Change dataset collected by surveillance cameras, our DECIDER similarly exceeds SOTA performance, where the main metric CIDEr is improved from 38.9 to 40.3, as illustrated in Table 1. The results demonstrate that DECIDER can be deployed in real-life monitoring scenarios.

Qualitative Results

We visualize several generated cases in Fig. 3. Cases (a) and (b) demonstrate that previous methods have incorrect perceptions of object orientation. When extended to natural scenes or photos with viewpoint changes, they are difficult to generate accurate descriptions that are consistent with the given image-pair, as illustrated in cases (c) and (e). Furthermore, case (d) shows that they fail to capture tiny changes in the surveillance perspective. In contrast, DECIDER can better perceive the tiny and position changes of objects.

Ablation Study and Analysis

We explore the effectiveness of the difference-aware cross-modal learning (*DCL.*), the "hard" positive (*Pos.*) and negative samples (*Neg.*), where we remove them as our baseline.

Effectiveness of Difference-aware Cross-modal Learning As shown in Table 3, the introduction of *DCL.* increases CIDEr/ROUGE-L by 1.9/0.5 and 0.7/0.7 than the baseline model on CLEVR-Change and Image-Editing-Request datasets, respectively. This is because the *DCL.* eliminates the redundant information during image-pair encoding, generating a compact yet effective visual difference representation for caption generation. Moreover, we visualize the difference representations without/with *DCL.* in Fig. 4, which illustrates that redundancy and irrelevant attention exist in the primary visual features and that *DCL.* can help our model more accurately locate the change regions of interest.

Effectiveness of Contrastive Diffusion Model Compared with the baseline model, we explicitly introduce "hard" positive and negative samples, bringing 1.8 and 0.8 CIDEr improvements on the two datasets, respectively. This superior performance gain proves that "hard" contrastive samples can

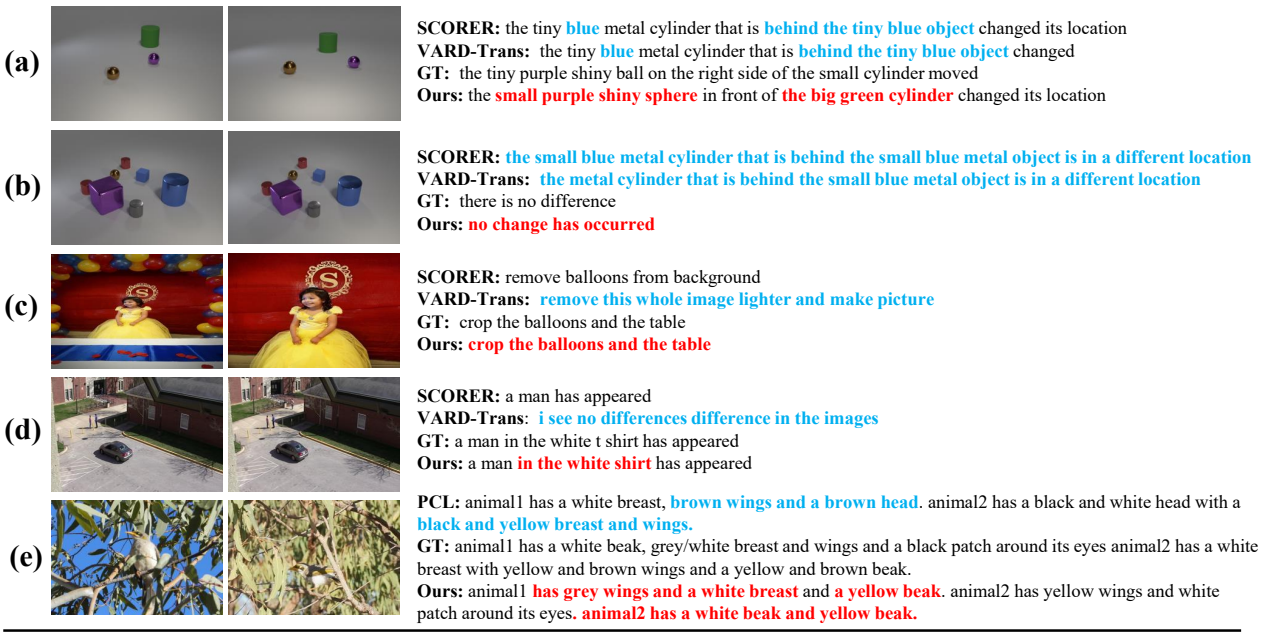


Figure 3: Visualization of generated captions. The first two rows illustrate two cases (object movement in (a) and no change in (b)) from CLEVR-Change. Case (c) from Image-Editing-Request describes the removal of entity under view-point change. Case (d) from Spot-to-Change presents a tiny difference in real scene. Case (e) from Birds-to-Words involves differences in bird appearance under complex backgrounds. The highlights in red/blue fonts indicate the correct/incorrect captions.

Dataset	Different Setting			Metrics	
	DCL.	*Pos.*	*Neg.*	C	R
CLEVR-Change	-	-	-	131.8	75.5
	✓	-	-	133.7	76.0
	-	✓	-	132.5	75.6
	-	-	✓	132.8	75.9
	-	✓	✓	133.6	76.2
	✓	✓	✓	**134.6**	**76.5**
Image-Editing-Request	-	-	-	38.3	41.1
	✓	-	-	39.0	41.8
	-	✓	-	38.6	41.4
	-	-	✓	38.7	41.6
	-	✓	✓	39.1	42.0
	✓	✓	✓	**39.5**	**42.4**

Table 3: Evaluation results of ablation studies on CLEVR-Change and Image-Editing-Request datasets.

facilitate our model to accurately understand and learn the cross-modal correspondence.

Dataset	λ	0	0.005	0.01	0.05	0.1
CLEVR-Change	**C**	134.1	134.4	**134.6**	134.2	133.7
	M	76.3	76.3	**76.5**	76.4	76.1

Table 4: Evaluation results of contrastive variational loss with different weights λ on CLEVR-Change.

Analysis of Contrastive Variational Loss We set different λ in Eq. (19) to observe the performance change as shown in Table 4. When λ is 0, *i.e.*, ignoring the optimization by negative samples, the performance is the worst. As λ increases from 0 to 0.01, the performance is improved since "hard" negative samples can provide discriminative information for our model to explore the correlation between visual changes and descriptions. However, the further increase of λ would lead to a performance drop because a too-large value overshadows the utility of positive samples.

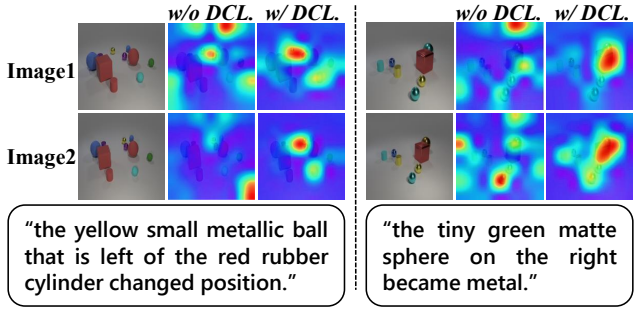


Figure 4: Visualization of visual difference representations with/without difference-aware cross-modal learning.

Conclusion

In this work, we propose a novel **D**ifference-aware **C**ontrastive **D**iffusion Model with **Ad**versarial **Per**turbations (DECIDER) for image change captioning. Extensive experiments confirm the effectiveness of the proposed DECIDER.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62272157).

References

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and Van Den Berg, R. 2021. Structured Denoising Diffusion Models in Discrete State-spaces. *Advances in Neural Information Processing Systems*, 34: 17981–17993.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Barber, D.; and Agakov, F. 2004. The im algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems*, 16(320): 201.

Chen, T.; Zhang, R.; and Hinton, G. 2022. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. In *International Conference on Learning Representations*.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, 1779–1788. PMLR.

Cheng, S.; and Sun, H. 2024. SPT: Sequence Prompt Transformer for Interactive Image Segmentation. arXiv:2412.10224.

Dieng, A. B.; Tran, D.; Ranganath, R.; Paisley, J.; and Blei, D. 2017. Variational Inference via χ Upper Bound Minimization. *Advances in Neural Information Processing Systems*, 30.

Forbes, M.; Kaeser-Chen, C.; Sharma, P.; and Belongie, S. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 708–717.

Gao, Z.; Guo, J.; Tan, X.; Zhu, Y.; Zhang, F.; Bian, J.; and Xu, L. 2022. Difformer: Empowering diffusion model on embedding space for text generation. *arXiv preprint arXiv:2212.09412*.

Gong, L.; Lin, Y.; Guo, S.; Lin, Y.; Wang, T.; Zheng, E.; Zhou, Z.; and Wan, H. 2023a. Contrastive pre-training with adversarial perturbations for check-in sequence representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4276–4283.

Gong, S.; Li, M.; Feng, J.; Wu, Z.; and Kong, L. 2023b. DiffuSeq-v2: Bridging Discrete and Continuous Text Spaces for Accelerated Seq2Seq Diffusion Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9868–9875.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10696–10706.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hoogeboom, E.; Nielsen, D.; Jaini, P.; Forré, P.; and Welling, M. 2021. Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions. *Advances in Neural Information Processing Systems*, 34: 12454–12465.

Huang, Q.; Liang, Y.; Wei, J.; Cai, Y.; Liang, H.; Leung, H.-f.; and Li, Q. 2021. Image Difference Captioning with Instance-level Fine-grained Feature Representation. *IEEE Transactions on Multimedia*, 24: 2004–2017.

Jhamtani, H.; and Berg-Kirkpatrick, T. 2018. Learning to Describe Differences Between Pairs of Similar Images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4024–4034.

Jiang, J.; Liu, Z.; and Zheng, N. 2023. Correlation information bottleneck: Towards adapting pretrained multimodal models for robust visual question answering. *International Journal of Computer Vision*, 1–23.

Kim, H.; Kim, J.; Lee, H.; Park, H.; and Kim, G. 2021. Agnostic Change Captioning with Cycle Consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2095–2104.

Lee, S.; Lee, D. B.; and Hwang, S. J. 2020. Contrastive Learning with Adversarial Perturbations for Conditional Text Generation. In *International Conference on Learning Representations*.

Li, S.; Li, W.; Cook, C.; Zhu, C.; and Gao, Y. 2018. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5457–5466.

Li, S.; Zhong, G.; Jin, Y.; Wu, X.; Zhu, P.; and Wang, Z. 2022a. A deceptive reviews detection method based on multidimensional feature construction and ensemble feature selection. *IEEE Transactions on Computational Social Systems*, 10(1): 153–165.

Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P. S.; and Hashimoto, T. B. 2022b. Diffusion-lm Improves Controllable Text Generation. *Advances in Neural Information Processing Systems*, 35: 4328–4343.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Lovelace, J.; Kishore, V.; Wan, C.; Shekhtman, E.; and Weinberger, K. Q. 2024. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36.

Luo, J.; Li, Y.; Pan, Y.; Yao, T.; Feng, J.; Chao, H.; and Mei, T. 2023. Semantic-conditional Diffusion Networks for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23359–23368.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust Change Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4624–4633.

Ruan, L.; Ma, Y.; Yang, H.; He, H.; Liu, B.; Fu, J.; Yuan, N. J.; Jin, Q.; and Guo, B. 2023. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10219–10228.

Savinov, N.; Chung, J.; Binkowski, M.; Elsen, E.; and van den Oord, A. 2021. Step-unrolled Denoising Autoencoders for Text Generation. In *International Conference on Learning Representations*.

Shi, X.; Yang, X.; Gu, J.; Joty, S.; and Cai, J. 2020. Finding It at Another Side: A Viewpoint-adapted Matching Encoder for Change Captioning. In *European Conference on Computer Vision*, 574–590. Springer.

Sun, H. 2024. Ultra-High Resolution Segmentation via Boundary-Enhanced Patch-Merging Transformer. arXiv:2412.10181.

Sun, H.; Xu, L.; Jin, S.; Luo, P.; Qian, C.; and Liu, W. 2024. PROGRAM: PROtotype GRAph Model based Pseudo-Label Learning for Test-Time Adaptation. In *The Twelfth International Conference on Learning Representations*.

Tan, H.; Dernoncourt, F.; Lin, Z.; Bui, T.; and Bansal, M. 2019. Expressing Visual Relationships via Language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1873–1883.

Tu, Y.; Li, L.; Su, L.; Du, J.; Lu, K.; and Huang, Q. 2023a. Adaptive Representation Disentanglement Network for Change Captioning. *IEEE Transactions on Image Processing*.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; and Huang, Q. 2024. SMART: Syntax-Calibrated Multi-Aspect Relation Transformer for Change Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2023b. Self-supervised Cross-view Representation Reconstruction for Change Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2805–2815.

Tu, Y.; Yao, T.; Li, L.; Lou, J.; Gao, S.; Yu, Z.; and Yan, C. 2021. Semantic Relation-aware Difference Representation Learning for Change Captioning. In *Findings of the Association for Computational Linguistics*, 63–73.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Yao, L.; Wang, W.; and Jin, Q. 2022. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3108–3116.

Yue, S.; Tu, Y.; Li, L.; Gao, S.; and Yu, Z. 2024. Multi-grained Representation Aggregating Transformer with Gating Cycle for Change Captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Zhong, G.; Ding, W.; Chen, L.; Wang, Y.; and Yu, Y.-F. 2023a. Multi-scale attention generative adversarial network for medical image enhancement. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(4): 1113–1125.

Zhong, G.; Guo, Y.; Yuan, J.; Zhang, Q.; Guan, W.; and Chen, L. 2024. PROMOTE: Prior-Guided Diffusion Model with Global-Local Contrastive Learning for Exemplar-Based Image Translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3313–3322.

Zhong, G.; Yuan, J.; Wang, P.; Yang, K.; Guan, W.; and Li, Z. 2023b. Contrast-augmented Diffusion Model with Fine-grained Sequence Alignment for Markup-to-Image Generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5311–5320.

Zhu, Y.; Wu, Y.; Olszewski, K.; Ren, J.; Tulyakov, S.; and Yan, Y. 2023. Discrete Contrastive Diffusion for Cross-Modal Music and Image Generation. In *International Conference on Learning Representations*.

Zhu, Z.; Wei, Y.; Wang, J.; Gan, Z.; Zhang, Z.; Wang, L.; Hua, G.; Wang, L.; Liu, Z.; and Hu, H. 2022. Exploring Discrete Diffusion Models for Image Captioning. *arXiv preprint arXiv:2211.11694*.